Statistical Inference of Optimal Allocations I: Regularities and their Implications

Kai Feng* Han Hong†

April 9, 2024

Abstract. In this paper, we develop a functional differentiability approach for solving statistical optimal allocation problems. We first derive Hadamard differentiability of the value function through a detailed analysis of the general properties of the sorting operator. Central to our framework are the concept of Hausdorff measure and the area and coarea integration formulas from geometric measure theory. Building on our Hadamard differentiability results, we demonstrate how the functional delta method can be used to directly derive the asymptotic properties of the value function process for binary constrained optimal allocation problems, as well as the two-step ROC curve estimator. Moreover, leveraging profound insights from geometric functional analysis on convex and local Lipschitz functionals, we obtain additional generic Fréchet differentiability results for the value functions of optimal allocation problems. These compelling findings motivate us to study carefully the first order approximation of the optimal social welfare. In this paper, we then present a double / debiased estimator for the value functions. Importantly, the conditions outlined in the Hadamard differentiability section validate the margin assumption from the statistical classification literature employing plug-in methods that justifies a faster convergence rate.

Keywords: Optimal treatment allocation, Functional differentiability and delta method, Plug-in classification, ROC curve, Double / debiased machine learning

JEL Codes: C14, C38, C44

*Institute of Economics, Tsinghua University. Email: fengk22@mails.tsinghua.edu.cn.
†Department of Economics, Stanford University. Email: doubleh@stanford.edu.

1 Introduction

Currently, as a consequence of the rapid digitization of social-economic data, characterizing individual heterogeneity has become crucial in the optimal allocation of scarce resources. Consider, for instance, the medical treatment assignment problem when doctors adopt different therapeutic approaches to different patients; the pretrial bail problem when judges decide where defendants will await trial; the college admission problem when committees assign students to various educational programs. In order to address the heterogeneity inherent in such problems, where the potential effects of a treatment on an outcome vary among individuals, personalized allocation is more of a necessity than just an option.

Based on past experimental or observational data, the core objective of a typical allocation problem is to find an allocation strategy that can perform well in the future. To achieve this goal, at least based on experience, an effective approach of decision making is to rely on information provided by prediction algorithms. (Kleinberg et al., 2018; Agrawal et al., 2019; Babina et al., 2024). In empirical applications, a common practice is to form a nonrandom treatment rule using an indicator function, in which a first step regression estimate of individual level treatment effect is plugged-in and benchmarked against certain decision thresholds. In this case, mathematically, optimal allocation is treated as a weighted classification problem.

A plug-in strategy is based on a population solution to the optimal allocation problem. For a joint distribution of $\{X, Y\}$, $Y = (Y_j, j = 0, \ldots, J)$, consider the *social welfare potential function*

$$\gamma(\lambda, g(\cdot)) := \max_{\phi \in \Phi} \mathbb{E}\left[\sum_j \lambda_j \phi_j(X) Y_j\right] = \max_{\phi \in \Phi} \mathbb{E}\left[\sum_j \lambda_j \phi_j(X) g_j(X)\right], \tag{1}$$

where Φ is the set of multi-critical functions, i.e. $\phi = (\phi_j, j = 0, \ldots, J)$, $\phi_j : \Omega \to R$, with (Ω, F, μ) being a Borel probability space, such that $\phi_j \in L^\infty(\mu), 0 \leqslant \phi_j \leqslant 1, \forall j \in \{0, \ldots, J\}$, and such that $\sum_{j=0}^J \phi_j(x) \equiv 1$ for all $x \in \Omega$; $g(\cdot) = \{g_j(\cdot), j = 0, \ldots, J\}$ and $g_j(x) = \mathbb{E}(Y_j | X = x)$. Under very general conditions, the problem posed in (1) admits an intuitive solution where the jth-category receives all allocation when $\mathbb{E}(\lambda_j Y_j | X)$ is uniquely the largest. In other words,

$$\phi_j(X) = \begin{cases} 1, & \text{if } \mathbb{E}(\lambda_j Y_j | X) > \mathbb{E}(\lambda_l Y_l | X),\ \forall l \neq j, \\ 0, & \text{if } \mathbb{E}(\lambda_j Y_j | X) < \mathbb{E}(\lambda_l Y_l | X),\ \exists l \neq j. \end{cases} \tag{2}$$

If there exists some l such that $\mathbb{E}(\lambda_j Y_j | X) = \mathbb{E}(\lambda_l Y_l | X)$, then $\phi(\cdot)$ can be divided among all the maximal indexes.

In this paper, we are interested in differentiating $\gamma(\lambda, g)$ functionally with respect to both λ and $g(\cdot)$. The solution of such problem can help answering the core problem of the statistical treatment rule literature, i.e. establishing probabilistic *regret* guarantees for treatment rules obtained from empirical data, as considered by Manski (2004) and the subsequent works. The regret there is defined as the difference between the social welfare potential and the expected utility achieved by an estimated decision rule. Our approach can also provide answers for multiple other problems of interdisciplinary interest. For instance, we will derive asymptotic distributions for the binary optimal allocation subject to one resource constraint and the receiver operating characteristic (ROC) curve.

Our paper is an attempt to synthesize the methodology and perspective of several strands of literature. Hereafter, the core nature of issues, as will be readily apparent, comes from the statistical treatment rule literature. A widely-known surprising phenomenon in the plug-in classification literature is that the second step classification is simpler than the first step regression, in the sense that the former usually converges faster than the latter (Devroye et al., 1996; Audibert and Tsybakov, 2007). Built on seminar works Chen et al. (2003) and Chernozhukov et al. (2018b) on method of moments allowing nonsmoothness, especially the latter, our paper can provide sets of conditions to support $\sqrt{n}$ and faster convergence rates of both the regret and other objective of interest. It will be shown that our differentiation results allow for asymptotic analysis under settings for both classical empirical processes and recent advances in double / debiased machine learning (Chernozhukov et al., 2018a, 2022).

To get some intuition of our approach, consider the finite dimensional derivatives of $\gamma(\lambda, g)$ with respect to λ when $J = 1$. Assume that $\mathbb{P}\{\lambda_1 g_1(X) = \lambda_0 g_0(X)\} = 0$. Then we can write,

$$\gamma(\lambda_0, \lambda_1, g) = \mathbb{E}\left[1\left(\lambda_0 g_0(X) > \lambda_1 g_1(X)\right)\lambda_0 g_0(X)\right] + \mathbb{E}\left[1\left(\lambda_1 g_1(X) > \lambda_0 g_0(X)\right)\lambda_1 g_1(X)\right]. \tag{3}$$

Formally, given the definition of $\gamma(\lambda_0, \lambda_1, g)$, we expect that

$$\begin{aligned}\frac{\partial \gamma(\lambda_0, \lambda_1, g)}{\partial \lambda_0} &= \mathbb{E}\left[1\left(\lambda_0 g_0(X) > \lambda_1 g_1(X)\right) g_0(X)\right], \\ \frac{\partial \gamma(\lambda_0, \lambda_1, g)}{\partial \lambda_1} &= \mathbb{E}\left[1\left(\lambda_1 g_1(X) > \lambda_0 g_0(X)\right) g_1(X)\right].\end{aligned}$$

A rigorous proof of the above is by direct calculation of the limit. It is not hard to see that the terms of major difficulties are

$$\begin{aligned}&\lim_{h\to 0}\frac{\mathbb{E}\left[1\left(\left(\lambda_0 + h\right) g_0(X) > \lambda_1 g_1(X)\right) - 1\left(\lambda_0 g_0(X) > \lambda_1 g_1(X)\right)\lambda_0 g_0(X)\right]}{h} \quad \text{and} \\ &\lim_{h\to 0}\frac{\mathbb{E}\left[1\left(\lambda_1 g_1(X) > \left(\lambda_0 + h\right) g_0(X)\right) - 1\left(\lambda_1 g_1(X) > \lambda_0 g_0(X)\right)\lambda_1 g_1(X)\right]}{h}.\end{aligned} \tag{4}$$

We will see later in section 3 that under mild conditions, the two terms in (4) both converge to the same integral on the level set (i.e. fiber) $\{x : \lambda_0 g_0(x) = \lambda_1 g_1(x)\}$ with opposing signs, so that they cancel out. As a consequence, the desired partial derivatives exist and take the displayed simple form.

Our approach in section 3 is greatly inspired by the *integration on manifold* approach in Chernozhukov et al. (2018b). In particular, we reapproach and generalize their results based on the concept of Hausdorff measure and powerful integration formulas, i.e. area and coarea formulas from geometric measure theory. By precise calculation, we manage to show that the Hadamard derivative of $\gamma(\lambda, g)$ contains terms that are integrals under certain Hausdorff measure. The Hadamard differentiability results are derived under mild primitive conditions, which differentiates us from previous works relying on differentibility or pathwise differentibility assumptions. See for example Sherman (1993) and Luedtke and van der Laan (2016a).

Returning to the general multivariate setting, the social welfare potential function in (1) is not only subadditive but also positive homogeneous of degree one with respect to both λ and $g(\cdot)$, respectively, which implies that $\gamma(\lambda, g(\cdot))$ is convex in $g(\cdot)$ given λ and is convex in λ given $g(\cdot)$. For example,

$$\begin{aligned}\gamma\left(\lambda, \alpha g + (1-\alpha) g'\right) &= \max_{\phi\in\Phi} \mathbb{E}\left[\sum_{j=0}^{J} \lambda_j \phi_j(X)\left(\alpha g_j(X) + (1-\alpha) g'_j(X)\right)\right] \\ &\leqslant \max_{\phi\in\Phi} \mathbb{E}\left[\sum_{j=0}^{J} \alpha\lambda_j\phi_j(X) g_j(X)\right] + \max_{\phi\in\Phi} \mathbb{E}\left[\sum_{j=0}^{J} (1-\alpha)\lambda_j\phi_j(X) g'_j(X)\right] \\ &= \alpha\gamma(\lambda, g) + (1-\alpha)\gamma\left(\lambda, g'\right).\end{aligned}$$

However, $\gamma(\lambda, g)$ is not necessarily jointly convex in (λ, g). We can not be assured that

$$\gamma\left(\alpha\lambda + (1-\alpha)\lambda', \alpha g + (1-\alpha) g'\right) \leqslant \alpha\gamma(\lambda, g) + (1-\alpha)\gamma\left(\lambda', g'\right).$$

Interestingly, the above simple observations seem to have gone completely unnoticed by the existing literature. In section 4, we apply deep results from geometric functional analysis, in particular those by Stegall (1978), Preiss (1990) and Lindenstrauss and Preiss (2003), to show that under nearly no assumptions, the social welfare potential function $\gamma(\lambda, g)$ is Fréchet differentiable on a generic set.

The surprising Fréchet differentiability property inpires us to reconsider the first order approximation of $\gamma(\lambda, g)$. Section 4 shows that the primitive conditions in section 3 can also be used to support the faster convergence rate in classical classification literature. See for example Audibert and Tsybakov (2007) among others. Further, the same methodology can be combined with recent double / debiased machine learning approach (Chernozhukov et al., 2018a, 2022) to speed up the estimation of $\gamma(\lambda, g)$. The roles of the regularity conditions clarified in this paper are fully demonstrated when they collaborate in tandem with the insights from the corresponding previous literature.

2 Literature

The construction of optimal decision rules from experimental or observational data have been considered by both the classical and recent literature in multidisciplines. Different terminology have been used, such as treatment allocation, policy learning in econometrics and individual treatment rule (ITR) in statistics. A large body of the literature, including Manski (2004), Qian and Murphy (2011), Zhao et al. (2012), Swaminathan and Joachims (2015), Zhou et al. (2017), Kitagawa and Tetenov (2018), Rai (2018), Luedtke and Chambaz (2020), Athey and Wager (2021), Mbakop and Tabord-Meehan (2021), Zhou et al. (2023) and Ben-Michael et al. (2024), among others, study in details this problem in an empirical risk minimization (ERM) framework. The basic objective there is to provide probabilistic bound for

$$R\left(\hat{\phi}, \phi^*\right) = \mathbb{E}_{X,Y}\left[\sum_j \phi_j^*\left(X\right) Y_j\right] - \mathbb{E}_{X,Y}\left[\sum_j \hat{\phi}_j\left(X\right) Y_j\right],$$
$$\text{for } \hat{\phi} \in \arg\max_{\phi\in\Phi_0} \sum_i \langle \phi\left(X_i\right), F\left(X_i, D_i, Y_i\right)\rangle,$$

where ϕ^* solves the population optimization program of (1). The function $F\left(X, D, Y\right)$ usually takes the inverse propensity weighting (IPW) or doubly robust (DR) form (see for example Kitagawa and Tetenov (2018) and Athey and Wager (2021)), and $\Phi_0 \subset \Phi$ is not too complex in an entropy sense.

The difficulty involved in the optimal allocation problem is due to the indicator function in the population solution which is hard to handle (Qian and Murphy, 2011). Our solution of such a problem built on a divergent literature motivated by very different purposes. To address the restoration of monotonicity in conditional quantile estimation, the functional derivative of the sorting operator in the univariate case is studied by Chernozhukov et al. (2010). The analysis of the sorting operator is then extended to the multivariate cases by Chernozhukov et al. (2018b) using calculus on manifold techniques. Kim and Pollard (1990) also hinted at a rudimentary form of calculus on manifold in the derivation of the large sample properties of cube root consistent estimators. Sasaki (2015) incorporated the tools based on the Hausdorff measure developed in fluid mechanics to characterize the information content of quantile partial derivatives for general structural functions.

The optimal allocation problem has also been discussed under a plethora of different but closely related settings. Examples include asymptotically minimax optimal decision under the limits of experiments framework (Hirano and Porter, 2009), optimal decision under minimax regret (Stoye, 2009; Tetenov, 2012; Ben-Michael et al., 2024), uniform confidence interval (Armstrong and Shen, 2023), allocation with spillover effects (Kitagawa and Wang, 2023a,b) and binary constrained optimal allocation (Bhattacharya and Dupas, 2012; Luedtke and van der Laan, 2016a). Our general functional Hadamard derivative results in subsection 3.3 are applicable for solving the last problem above. This problem is also closely relevant to the receiver operating characteristic (ROC) curve widely used in biostatistics and computer science. A previous literature that has studied the inference of the ROC curves, including Hsieh and Turnbull (1996), Lloyd (1998), Li et al. (1999), Hall et al. (2004) and Bertail et al. (2008), did not account for the estimation error in the sample propensity score. An exception is Luckett et al. (2021), where the authors derive the uniform asymptotic results for estimated ROC curve when the classification is done by a support vector machine (SVM). Unlike the studies above, we derive the asymptotic properties of the ROC curves obtained by plugging-in the propensity score estimators, which is consistent with the usage of predictive classifiers in various disciplines.

The methodology in this paper also has intriguing connection with the faster convergence phenomenon in the classification literature (Devroye et al., 1996; Audibert and Tsybakov, 2007). Consider, especially, the margin assumption (MA) evoked by works including Tsybakov (2004), Audibert and Tsybakov (2007) and Boucheron et al. (2005), i.e.

$$\mathbb{P}\left\{\left|p\left(X\right)-c\right|<t\right\}\leqslant Ct^{\alpha}$$

for fixed c and some constant $C>0$ and $\alpha\geqslant 0$. Our conditions in section 3 directly imply the marginal assumption with $\alpha\geqslant 1$. This observation can be further combined with the recent double / debiased machine learning approach (Chernozhukov et al., 2018a, 2022) to form a relatively primitive level, plug-in and non empirical process complement of the optimal allocation literature.

3 Hadamard differentiability

Following the seminal research of Chernozhukov et al. (2018b), we consider (3) as a functional mapping from $(\lambda, g(\cdot))$ to $\gamma(\lambda, g(\cdot))$, and are interested in the differential change in $\gamma(\lambda, g(\cdot))$ induced by a marginal change in $(\lambda, g(\cdot))$. More precisely, we rigorously derive the Hadamard derivative of $\gamma(\lambda, g(\cdot))$ with respect to $(\lambda, g(\cdot))$, which justifies and generalizes the calculations in (4). In this section, we assume that the underlying distribution of X, denoted as μ, is absolutely continuous with respect to $\mathcal{L}_n$, the n-dimensional Lebesgue measure. Then we write

$$\mathbb{E}\left[\sum_{j=0}^{J}\lambda_j\phi_j\left(X\right)g_j\left(X\right)\right]=\sum_{j=0}^{J}\int\lambda_j\phi_j\left(x\right)g_j\left(x\right)d\mu\left(x\right)=\sum_{j=0}^{J}\int\lambda_j\phi_j\left(x\right)g_j\left(x\right)f\left(x\right)d\mathcal{L}_nx.$$

where $f(x)$ is the density function. Typically, the optimizing policy function takes the form of $\phi_j(x)=\prod_{l\neq j}1\left(\lambda_j g_j(x)>\lambda_l g_l(x)\right)$. For simplicity, if we abbreviate $\lambda_j g_j(x)$ as $g_j(x)$, then the main object that we are going turns out to be

$$\int g_j\left(x\right)f\left(x\right)\prod_{l\neq j}1\left(g_j\left(x\right)>g_l\left(x\right)\right)d\mathcal{L}_nx$$

which can be further generalized to a form of the following sorting operator (as defined by Chernozhukov et al. (2018b)),

$$\int g_j(x) f(x) \prod_l 1(h_l(x) > c) \, d\mathcal{L}_n x, \tag{5}$$

where the number of terms in the product should be less than n, the dimension of X.

3.1 Hausdorff measure and integration formulas

The following discussions are based on standard references by Federer (1969) and Evans and Garzepy (2015). In order to rigorously present surface integral in Chernozhukov et al. (2018b), this section first aims to define extrinsically the intuitive notion of length, area, and volume. For this purpose, the Carathéodory criterion is used to construct a measure space from a σ-subadditive set function (outer measure) on power set of $\mathbb{R}^N$. More specifically, we introduce Hausdorff's construction, which does not require any local parameterization on these sets. Therefore no regularity assumption is needed.

Definition 3.1.1. (Outer measure) For a set O, an outer measure is a function

$$\mu^* : 2^O \to [0, \infty] \quad \text{such that}$$

(a) $\mu^*(\varnothing) = 0$.

(b) For arbitrary subsets A, B, $A \subset B \subset O$, $\mu^*(A) \leqslant \mu^*(B)$.

(c) For arbitrary subsets $B_1, B_2, \ldots$ of O,

$$\mu^*\left(\bigcup_{i=1}^{\infty} B_i\right) \leqslant \sum_{i=1}^{\infty} \mu^*(B_i).$$

If an outer measure μ^* is defined on a metric space (O, d) and satisfies

$$d(A, B) > 0 \Rightarrow \mu^*(A \cup B) = \mu^*(A) + \mu^*(B), \quad \forall A, B \subset O.$$

Then μ^* is called a metric outer measure.

In this paper, we use a series of metric outer measure with the following specific construction, called Hausdorff outer measure.

Definition 3.1.2. (Hausdorff outer measure) Let d be the metric on O. For arbitrary subset

$E \subset O$, define its diameter as $\operatorname{diam} E = \sup_{x_1, x_2 \in E} d(x_1, x_2)$, $\operatorname{diam} \varnothing = 0$. Let

$$\alpha_k = \frac{\pi^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}+1\right)}, \quad k \in \mathbb{N} = \{0, 1, \ldots\},$$

where $\Gamma(\cdot)$ is the gamma function. For arbitrary $E \subset O$, and for $\delta > 0$, define

$$\mathcal{H}^*_{k,\delta}(E) = \inf \left\{ \sum_{j \geqslant 1} \alpha_k \left(\frac{\operatorname{diam} B_j}{2} \right)^k : E \subset \bigcup_{j \geqslant 1} B_j, \operatorname{diam} B_j \leqslant \delta \right\}.$$

The k-dimensional Hausdorff outer measure of E is defined as $\mathcal{H}^*_k(E) = \lim_{\delta \downarrow 0} \mathcal{H}^*_{k,\delta}(E)$, and can be verified to be a metric outer measure on the metric space (O, d).

Definition 3.1.3. (Hausdorff measure) Let $k \in \mathbb{N}$, $\mathcal{M}_k = \mathcal{M}_k(O)$ is the σ-algebra generated by $\mathcal{H}^*_k$ by the Carathéodory criterion (See Theorem T.1.1). Then the restriction of $\mathcal{H}^*_k$ to $\mathcal{M}_k$, $\mathcal{H}_k = \mathcal{H}^*_k|_{\mathcal{M}_k}$ is called Hausdorff measure on $\mathcal{M}_k$.

In $\mathbb{R}^k$, $\mathcal{H}^*_k = \mathcal{L}^*_k$, $\mathcal{H}_k = \mathcal{L}_k$, where $\mathcal{L}^*_k$ denotes the Lebesgue outer measure and $\mathcal{L}_k$ denotes the Lebesgue measure on $\mathbb{R}^k$. $\mathcal{H}^*_0(E)$ measures the cardinality of the set E:

$$\mathcal{H}^*_0(E) = \mathcal{H}_0(E) = \begin{cases} \text{number of element in } E, & E \text{ is empty or finite,} \\ \\ +\infty, & E \text{ is infinite.} \end{cases}$$

In this application, we focus on the Hausdorff measure defined on Euclidean spaces $\mathbb{R}^k$.

Definition 3.1.4. (Jacobian) For $k \in \{1, 2, \ldots, n\}$, define the Jacobian at x as

$$Jf(x) = \begin{cases} \sqrt{\det\left(\nabla f(x)^T \nabla f(x)\right)}, & f : E \subset \mathbb{R}^k \mapsto \mathbb{R}^n, \\ \sqrt{\det\left(\nabla f(x) \nabla f(x)^T\right)}, & f : E \subset \mathbb{R}^n \mapsto \mathbb{R}^k. \end{cases}$$

The following Theorem 3.1.1 for change of variable generalizes the conventional integration change of variable formula for invertible mapping to allow for non-invertible mapping and an increase in dimension.

Theorem 3.1.1. *Area formula*
Let $E \subset \mathbb{R}^k$ be an open set, $k \in \{1, 2, \ldots, n\}$, $\psi : E \mapsto \mathbb{R}^n$ be a C^1 or Lipschitz function. Then for

all measurable $S \subset E$,

$$\int_S J\psi(x)\, d\mathcal{L}_k x = \int_{\psi(S)} \mathcal{H}_0\left(S \cap \psi^{-1}(y)\right) d\mathcal{H}_k y. \tag{6}$$

If a measurable function $f : S \mapsto \mathbb{R}$ is nonnegative or if the left hand side of (7) is finite, then the following equality holds,

$$\int_S f(x)\, J\psi(x)\, d\mathcal{L}_k x = \int_{\psi(S)} \left[\int_{S\cap\psi^{-1}(y)} f(x)\, d\mathcal{H}_0 x\right] d\mathcal{H}_k y. \tag{7}$$

Furthermore, the next *curving Fubini-Tonelli* theorem allows for slicing among curving surfaces and mapping from a higher dimension to a lower dimension.

Theorem 3.1.2. *Coarea formula*
Let $E \subset \mathbb{R}^n$ be an open set, $k \in \{1, 2, \ldots, n\}$, $\varphi : E \mapsto \mathbb{R}^k$ be a C^1 or Lipschitz function, then for all measurable $S \subset E$,

$$\int_S J\varphi(x)\, d\mathcal{L}_n x = \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y. \tag{8}$$

If a measurable function $f : S \mapsto \mathbb{R}$ is nonnegative or if the left hand side of (9) is finite, then the following equality holds,

$$\int_S f(x)\, J\varphi(x)\, d\mathcal{L}_n x = \int_{\mathbb{R}^k} \left[\int_{S\cap\varphi^{-1}(y)} f(x)\, d\mathcal{H}_{n-k} x\right] d\mathcal{H}_k y. \tag{9}$$

Remark 3.1.1. (7) and (9) are direct corollaries of (6) and (8), respectively, by a standard simple function approximation of measurable functions. (7), (9) and the classical $\mathbb{R}^n \mapsto \mathbb{R}^n$ change of variables formula, can be collectively written in a unified formula:

$$\int_S f(x)\, J\varphi(x)\, d\mathcal{L}_n x = \int_{\mathbb{R}^m} \left[\int_{S\cap\varphi^{-1}(y)} f(x)\, d\mathcal{H}_{\max\{n-m,0\}} x\right] d\mathcal{H}_{\min\{n,m\}} y.$$

To the best of our knowledge, a result similar to Theorem 3.1.1 was first published in Federer (1944). The coarea formula in the form of (9) was first published in Federer (1959), one of the most important papers by Federer. Results similar to Theorem 3.1.2 were presented earlier by Aleksandr Semyonovich Kronrod, Herbert Federer, Laurence Chisholm Young (L.C.Young), Ennio De Giorgi and maybe even others. The statements of the area and coarea formulas in Theorems 3.1.1 and 3.1.2 are closer to Federer (1969) and the more readable Evans and Garzepy (2015). Roughly speaking, generalizations of both Theorem 3.1.1 and Theorem 3.1.2 can still work when the Euclidean spaces therein are replaced by certain kind of *surface*, e.g. Riemannian manifolds and rectifiable sets of Euclidean spaces. See 3.2.20-3.2.22 and 3.2.46 in Federer (1969) for classical developments. In technical addendum T, we provide readable proofs to both Theorem 3.1.1 and Theorem 3.1.2 together with a complete set of preliminary results needed in their derivation.

Definition 3.1.5. (Critical and regular point and value) Let $n, m \in \{1, 2, \ldots\}$, $E \subset \mathbb{R}^n$ be an open set, $f : E \mapsto \mathbb{R}^m$ be a C^1 function. A point $x \in E$ is called a critical point if $Jf(x) = 0$; otherwise x is called a regular point. If $c = f(x)$ for some critical point x, then c is called a critical value; otherwise c is called a regular value. For simplicity, in the main text we only consider regular value $c \in f(E)$.

The area and coarea formulas in Theorems 3.1.1 and 3.1.2 are closely related to another important result called Morse-Sard Theorem, which essentially states that the set of critical values is a null set. See for example Theorem T.1.10. Corollary T.3.2, which follows directly from Theorem 3.1.1, is indeed Theorem T.1.10 when $n = m$. In economics, one of the most well-known applications of the Sard type theorem is Debreu (1970) which established that except for a null set of *economies*, every *economy* has a finite set of equilibria. For a relatively simple derivation of the Morse-Sard theorem, see Figalli (2008). The basic ideas in Figalli (2008) is to make use of the Morrey inequality, which is then combined with Whitney extension theorem T.1.3 to prove the Morse-Sard theorem in a stronger Sobolev sense, which implies the C^r case. Evans and Garzepy (2015) shows in detail how to use the coarea formula (9) to develop the Morrey inequality.

3.2 Hadamard derivatives of the sorting operator

In this section, we introduce our mathematics results on the Hadamard differentiability of the sorting operator in (5), where we intend to compute

$$\lim_{t\to 0}\frac{1}{t}\left[\int\left(g_j(x)+tG_j(x)\right)f(x)\,1\left(h(x)+tH(x)>c\right)d\mathcal{L}_n x-\int g_j(x)\,f(x)\,1\left(h(x)>c\right)d\mathcal{L}_n x\right].$$

It should be clear that the main term that needs to be accounted for is

$$\lim_{t\to 0}\frac{1}{t}\left[\int g_j(x)\,f(x)\,1\left(h(x)+tH(x)>c\right)d\mathcal{L}_n x-\int g_j(x)\,f(x)\,1\left(h(x)>c\right)d\mathcal{L}_n x\right].$$

Therefore we focus on the following simplified form of the sorting operator (with respect to the Lebesgue measure):

$$F(h,c) := \int_{h(x)>c} f(x)\,d\mathcal{L}_n x = \int \mathbb{1}\left(h(x)>c\right)f(x)\,d\mathcal{L}_n x, \tag{10}$$

where, without loss of generality, the $g_j(x)\,f(x)$ term is replaced by $f(x)$ above.

First, consider a special case of the sorting operator where $h(x)$ is a scalar function, where the directions of differentiation are limited to 1. Let h be a C^1 or Lipschitz function, with $\|\nabla h(x)\| > 0, \mathcal{L}_n$ *a.e.* x. Suppose also that f is a integrable (nonnegative measurable) function. Then by the coarea formula Theorem 3.1.2,

$$\int_{h(x)>c} f(x)\, d\mathcal{L}_n x = \int_c^\infty \left[\int_{h^{-1}(c')} \frac{f(x)}{\|\nabla h(x)\|} d\mathcal{H}_{n-1} x\right] d\mathcal{L}_1 c'.$$

Further, if $\operatorname{ess\,inf} \|\nabla h(x)\| > 0$, and f is integrable, then

$$\frac{d}{dc}\left(\int_{h(x)>c} f(x)\, d\mathcal{L}_n x\right) = -\int_{h^{-1}(c)} \frac{f(x)}{\|\nabla h(x)\|} d\mathcal{H}_{n-1} x$$

for $\mathcal{L}_1$ *a.e.* c, by the Lebesgue differentiation theorem, where the derivative is conventionally defined. The right hand side above is also a derivative in the distributional (weak) sense.

Definition 3.2.1. (Hadamard differentiability) Let $\mathcal{X}$ and $\mathcal{Y}$ be normed spaces equipped with norms $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$. Consider the map $\phi : \mathcal{X}_\phi \subset \mathcal{X} \to \mathcal{Y}$. Then ϕ is called Hadamard differentiable at $\theta \in \mathcal{X}_\phi$ tangentially to $\mathcal{X}_0 \subset \mathcal{X}$, if there is a continuous linear map $\phi'_\theta : \mathcal{X}_0 \to \mathcal{Y}$ such that:

$$\left\|\frac{\phi(\theta + t_n x_n) - \phi(\theta)}{t_n} - \phi'_\theta(x)\right\|_{\mathcal{Y}} \to 0, \text{ for all } t_n \to 0 \text{ and } x_n \to x, \tag{11}$$

as $n \to \infty$, where $\{x_n\} \subset \mathcal{X}$, $x \in \mathcal{X}_0$, and $\theta + t_n x_n \in \mathcal{X}_\phi$ for n large enough.

Remark 3.2.1.

1. When $\mathcal{X}_0$ is a linear subspace, by Fang and Santos (2019) Proposition 2.1, $t_n \to 0$ can be replaced by by $t_n \downarrow 0$ in the definition of Hadamard differentiability without loss of generality.
2. The derivative ϕ'_h that satisfies equation (11) is necessarily positively homogeneous of degree one but not necessarily continuous or linear. See Shapiro (1990).

To verify Hadamard differentiability and the corresponding derivative of the sorting operator, we need to calculate

$$\lim_{n\to\infty} \frac{1}{t_n} \int \left[\mathbb{1}(h(x) + t_n H_n(x) > c) - \mathbb{1}(h(x) > c)\right] f(x)\, d\mathcal{L}_n x,$$

where $H_n \to H$ and $t_n \downarrow 0$ as $n \to \infty$. Intuitively, this limit should be closely related to a type of integral along the *surface* $\{x : h(x) = c\}$. The concept of Hausdorff measure $\mathcal{H}_k$ (see definition 3.1.3) essentially defines *area of surface* of k-dimensional volume in n-dimensional space, $k \in \{1, \ldots, n\}$. The important insight of Chernozhukov et al. (2018b) is represented by the integral under $\mathcal{H}_{n-1}$.

Let $E \subset \mathbb{R}^n$ be an open set, $k \in \{1, 2, \ldots, n\}$, $h : E \mapsto \mathbb{R}^k$ be a C^1 function, $f : E \mapsto \mathbb{R}$ be a continuous function. Let $C(E)$ denote the space of continuous functions on E equipped with the sup-norm, and $C_b(E)$ that of bounded continuous functions. We need the following technical assumptions about h and f.

Assumption 3.2.1. *The function f is continuous with compact support $K_f \subset E$.*

Assumption 3.2.2. *The support K_f consists of only regular points of the function h.*

Assumption 3.2.1 guarantees that $K_f \cap \{x : h(x) = c'\}$ is close to $K_f \cap \{x : h(x) = c\}$ when c' is close to c. It can be replaced by alternative conditions, such as those in Assumption 3.2.1$'$. The purpose of the Assumption 3.2.2 is to guarantee that h is regular in a neighborhood of K_f in order to apply the implicit function theorem.

Assumption 3.2.1$'$. *For all $c \in h(E)$, there exists a neighborhood $\mathcal{N}_c$ of c such that*

$$\overline{\bigcup_{c' \in \mathcal{N}_c} h^{-1}(c')} \subset E$$

and is bounded. The function f is continuous.

Theorem 3.2.1. *Let $h : E \to \mathbb{R}^k$ be a C^1 function, E an open subset of $\mathbb{R}^n$, $k \in \{1, 2, \ldots, n\}$. Consider (10). Let Assumption 3.2.1 (or 3.2.1$'$) and Assumption 3.2.2 hold. Let $\mathcal{D} \subset h(E)$ be a compact subset. Then, the map $F(h, c) : C(E) \to \mathbb{R}$ is Hadamard differentiable uniformly on $\mathcal{D}$ at h tangentially to $C(E)$ ($C_b(E)$ under Assumption 3.2.1$'$). The derivative is given by*

$$F'_{h,c}(H, 0) = \sum_i \int_{y > \tau_{\neg i}(c)} \left[\int_{h^{-1}(c'[i])} \frac{H_i(x) f(x)}{Jh(x)} d\mathcal{H}_{n-k} x \right] d\mathcal{L}_{k-1} y,$$

where $\tau_{\neg i}(c)$ denotes the $\mathbb{R}^k \to \mathbb{R}^{k-1}$ coordinate projection except the i-th coordinate, $c'[i]$ is the k-dimensional vector such that $\tau_{\neg i}(c'[i]) = y$ and $\tau_i(c'[i]) = c_i$.

The most common application of Theorem 3.2.1 is when $k = 1$, which is listed separately below as Theorem 3.2.2 with even weaker conditions.

Assumption 3.2.2$'$. *$\exists c \in h(E)$ such that $K_f \cap h^{-1}(c)$ consists of only regular points of h.*

Theorem 3.2.2. *Consider when $k = 1$, i.e. when h is a scalar-valued function. Let Assumption 3.2.1 (or 3.2.1$'$) and Assumption 3.2.2$'$ hold. Then, there exists a bounded closed interval $\mathcal{D} \subset$*

$h(E)$, $c \in \mathcal{D}$, such that the map $F(h,c) : C(E) \to \mathbb{R}$ is Hadamard differentiable uniformly on $\mathcal{D}$ at h tangentially to $C(E)$ ($C_b(E)$ under Assumption 3.2.1′). The derivative is given by

$$F'_{h,c}(H,0) = \int_{h^{-1}(c)} \frac{H(x) f(x)}{\|\nabla h(x)\|} d\mathcal{H}_{n-1}x. \tag{12}$$

Embedded in the proof of Theorem 3.2.2 is the continuity of the derivative of the sorting operator, expressed in the following lemma, which also handle the case when $k > 1$.

Proposition 3.2.3. *Under the conditions in Theorem 3.2.1 or Theorem 3.2.2*

$$\int_{h^{-1}(c)} \frac{f(x)}{Jh(x)} d\mathcal{H}_{n-k}x$$

is continuous on $h(E)$ or $\mathcal{D}$.

Proofs of Theorems 3.2.1, Theorem 3.2.2 and Proposition 3.2.3. We will prove these three results in several parts. Before proceeding to the more general case of $k > 1$ in Theorem 3.2.1, we first prove a pointwise version of Theorem 3.2.2 for the special case when $k = 1$. Proposition 3.2.3 can be proved by a similar derivation in passing. We finalize by demonstrating the uniformity of convergence. The following proof is derived under Assumption 3.2.1.

part 1 First consider Theorem 3.2.2.
step 1 We claim that there exists an $\eta > 0$ small enough such that for all c' with $|c' - c| < \eta$, we can get change of variable formulas simultaneously. Consider

$$\Psi(x'_1, \ldots, x'_n, c') := h(x'_1, \ldots, x'_n) - c'.$$

Without loss of generality, we may assume that $\nabla_{x_1}\Psi(x_1, \ldots, x_n, c)$ is full rank. Then by the implicit function theorem (see for example Theorem T.1.9), there exists an open set $B_{x_1} \times B_{x_2,\ldots,x_n,c} \subset U \subset \mathbb{R}^{n+1}$, where U is a neighborhood of $(x_1, \ldots, x_n, c)$, such that for some positive vectors α, β,

$$\begin{aligned} B_{x_1} &= \left\{x'_1 \in \mathbb{R} : |x'_1 - x_1| < \alpha\right\} \\ B_{x_2,\ldots,x_n,c} &= \left\{(x'_2, \ldots, x'_n, c') \in \mathbb{R}^n : |(x'_2, \ldots, x'_n, c') - (x_2, \ldots, x_n, c)| < \beta\right\}, \end{aligned}$$

and a C^1 implicit function $\xi(\cdot)$ defined on $B_{x_2,\ldots,x_n,c}$ such that

$$\Psi(x'_1, \ldots, x'_n, c') = 0 \Leftrightarrow (x'_1) = \xi(x'_2, \ldots, x'_n, c')$$

for all $(x'_1, \ldots, x'_n, c') \in B_{x_1} \times B_{x_2,\ldots,x_n,c}$. For all $x \in h^{-1}(c) \cap K_f$, we can find intervals like $B_{x_1} \times B_{x_2,\ldots,x_n,c}$ in the above. The implicit function theorem may not be always about the first dimension. However, an implicit function for some $x'_m, m \in \{1, \ldots, n\}$ always exists by the regular point assumption 3.2.2′. By the compactness of $h^{-1}(c) \cap K_f$, there exists a finite open cover denoted as $\{B_j\} \coloneqq \left\{B_{x_{j,1}} \times B_{x_{j,2,\ldots,x_{j,n}}}\right\}$ of $h^{-1}(c) \cap K_f$. Now, we claim that there exists $\eta > 0$, such that for all c' satisfying $|c' - c| < \eta$, $h^{-1}(c') \cap K_f \subset \bigcup_j B_j$. This claim is proven by contradiction. Suppose $\bigcup_j B_j$ will not cover $h^{-1}(c') \cap K_f$ for some $c' \neq c$, such that $|c' - c| < \eta$ where η is arbitrarily chosen. In other words, $\forall \eta > 0$, $\exists c'$, $|c' - c| < \eta$, such that $h^{-1}(c') \cap K_f \not\subset \bigcup_j B_j$. Then there exists a sequence $\{(x_i, c_i)\}_{i=1}^{\infty}$ such that

$$x_i \in h^{-1}(c_i) \cap K_f, x_i \notin \bigcup_j B_j.$$

By the compactness of $h^{-1}([a,b]) \cap K_f$ and the Bolzano-Weierstrass theorem, there exists a convergent subsequence $\{x_{i_l}\}_{l=1}^{\infty}$ such that $\lim_{l\to\infty} h(x_{i_l}) = c$ and thus $\lim_{l\to\infty} x_{i_l} \in h^{-1}(c) \subset \bigcup_j B_j$. Therefore, $x_{i_l} \in \bigcup_j B_j$ for all sufficiently large l, which is a contradiction.

step 2 To calculate the change of variables formula analytically, consider

$$\psi_{B_j}\left(x'_{j,2}, \ldots, x'_{j,n}, c'\right) = \left(\xi_{B_j}\left(x'_{j,2}, \ldots, x'_{j,n}, c'\right), x'_{j,2}, \ldots, x'_{j,n}\right)^T,$$

where $\left(x'_{j,2}, \ldots, x'_{j,n}\right)$ are used by the implicit function theorem to obtain the local explicit function $\xi_{B_j}(\cdot)$ for $x'_{j,1}$. Then, by the generalized matrix determinant lemma (for clarity the subscript j is omitted), as used by Chernozhukov et al. (2018b),

$$\begin{aligned}
J\psi_{B_j}\left(x'_{2,\ldots,n}, c'\right) &= \det\left(\left(\left(\left[\frac{\partial h}{\partial x_1}\right]^{-1} \nabla_{x_{2,\ldots,n}} h\right)^T, I_{n-1}\right)\begin{pmatrix}\left[\frac{\partial h}{\partial x_1}\right]^{-1} \nabla_{x_{2,\ldots,n}} h \\ I_{n-1}\end{pmatrix}\right)^{\frac{1}{2}} \\
&= \det\left(I_{n-1} + \left(\left[\frac{\partial h}{\partial x_1}\right]^{-1} \nabla_{x_{2,\ldots,n}} h\right)^T \left(\left[\frac{\partial h}{\partial x_1}\right]^{-1} \nabla_{x_{2,\ldots,n}} h\right)\right)^{\frac{1}{2}} \\
&= \left(\det\left(I_{n-1}\right) \det\left(I_1 + \left(\left[\frac{\partial h}{\partial x_1}\right]^{-1} \nabla_{x_{2,\ldots,n}} h\right) I_{n-1} \left(\left[\frac{\partial h}{\partial x_1}\right]^{-1} \nabla_{x_{2,\ldots,n}} h\right)^T\right)\right)^{\frac{1}{2}} \\
&= \left(\det\left(\left[\frac{\partial h}{\partial x_1}\right]\left[\frac{\partial h}{\partial x_1}\right]^T + \left(\nabla_{x_{2,\ldots,n}} h\right)\left(\nabla_{x_{2,\ldots,n}} h\right)^T\right)\right)^{\frac{1}{2}} \left|\det\left(\left[\frac{\partial h}{\partial x_1}\right]^{-1}\right)\right| \\
&= \left(\det\left((\nabla h)(\nabla h)^T\right)\right)^{\frac{1}{2}} \left|\det\left(\left[\frac{\partial h}{\partial x_1}\right]^{-1}\right)\right| \\
&= Jh \left|\det\left(\left[\frac{\partial h}{\partial x_1}\right]^{-1}\right)\right|.
\end{aligned}$$

In the above, for simplicity, we omit the point at which the derivatives are calculated, where

$$\frac{\partial h}{\partial x_1} = \frac{\partial}{\partial x_1} h\left(\xi_{B_j}\left(x'_{2,\ldots,n}, c'\right), x'_2, \ldots, x'_n\right) \text{ and}$$
$$\nabla_{x_{2,\ldots,n}} h = \nabla_{x_{2,\ldots,n}} h\left(\xi_{B_j}\left(x'_{2,\ldots,n}, c'\right), x'_2, \ldots, x'_n\right).$$

step 3 Consider a sequence $t_n \downarrow 0$. First let $H_n \equiv H$ for all n. By definition of Hadamard differentiability, we need to calculate

$$\lim_{n\to\infty} \frac{1}{t_n}\left[\int 1\left(h(x) + t_n H(x) > c\right) f(x)\, d\mathcal{L}_n x - \int 1\left(h(x) > c\right) f(x)\, d\mathcal{L}_n x\right].$$

By coarea formula Theorem 3.1.2,

$$\begin{aligned}
&\frac{1}{t_n}\left[\int 1\left(h(x) + t_n H(x) > c\right) f(x)\, d\mathcal{L}_n x - \int 1\left(h(x) > c\right) f(x)\, d\mathcal{L}_n x\right] \\
&= \frac{1}{t_n}\int\left[\int_{h^{-1}(c')\cap K_f} \frac{\left(1\left(c' + t_n H(x) > c\right) - 1\left(c' > c\right)\right) f(x)}{Jh(x)} d\mathcal{H}_{n-1} x\right] d\mathcal{L}_1 c' \\
&= \int_{c - t_n M}^{c + t_n M}\left[\frac{1}{t_n}\int_{h^{-1}(c')\cap K_f} \frac{\left(1\left(c' + t_n H(x) > c\right) - 1\left(c' > c\right)\right) f(x)}{Jh(x)} d\mathcal{H}_{n-1} x\right] d\mathcal{L}_1 c',
\end{aligned}$$

where $M = \max_{x\in K_f}|H(x)| < +\infty$. Then, we can apply the area formula Theorem 3.1.1 to calculate, for each j,

$$\begin{aligned}
&\frac{1}{t_n}\int\left[\int_{h^{-1}(c')\cap B_j\cap K_f} \frac{\left[1\left(c' + t_n H(x) > c\right) - 1\left(c' > c\right)\right] f(x)}{Jh(x)} d\mathcal{H}_{n-1} x\right] d\mathcal{L}_1 c' \\
&= \frac{1}{t_n}\int\left[\int_{B_{x_{j,2},\ldots,x_{j,n}}} \frac{\left[1\left(c' + t_n H\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right) > c\right) - 1\left(c' > c\right)\right] f\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right)}{\left|\det\left(\frac{\partial}{\partial x_1} h\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right)\right)\right|} d\mathcal{L}_{n-1} x\right] d\mathcal{L}_1 c'.
\end{aligned}$$

Next by the Fubini-Tonelli theorem, the above is equal to

$$\int_{B_{x_{j,2},\ldots,x_{j,n}}}\left[\frac{1}{t_n}\int \frac{\left[1\left(c' + t_n H\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right) > c\right) - 1\left(c' > c\right)\right] f\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right)}{\left|\det\left(\frac{\partial}{\partial x_1} h\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right)\right)\right|} d\mathcal{L}_1 c'\right] d\mathcal{L}_{n-1} x.$$

Since

$$\begin{aligned}
&\frac{1}{t_n}\int\left[1\left(c' + t_n H\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right) > c\right) - 1\left(c' > c\right)\right] d\mathcal{L}_1 c' \\
&= \frac{1}{t_n}\int_{c - t_n M}^{c + t_n M}\left[1\left(c' + t_n H\left(\psi_{B_j}\left(x_{2,\ldots,n}, c'\right)\right) > c\right) - 1\left(c' > c\right)\right] d\mathcal{L}_1 c' \to H\left(\psi_{B_j}\left(x_{2,\ldots,n}, c\right)\right),
\end{aligned}$$

by the dominated convergence theorem (DCT) and with the help of the partition of unity theorem

T.1.11 (as in Chernozhukov et al. (2018b)), and by the area formula again, we get

$$\lim_{n\to\infty}\frac{1}{t_n}\int\left[1\left(h\left(x\right)+t_nH\left(x\right)>c\right)-1\left(h\left(x\right)>c\right)\right]f\left(x\right)d\mathcal{L}_nx=\int_{h^{-1}(c)}\frac{H\left(x\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x.$$

step 4 Now consider functions $H\left(x\right)+\zeta$, for arbitrary $\zeta>0$. Since by assumption $\sup_{x\in E}\left|H_n\left(x\right)-H\left(x\right)\right|\to 0$, for all $\zeta>0$, for sufficiently large n, $H\left(x\right)+\zeta>H_n\left(x\right)$ for all x. Therefore,

$$\begin{aligned}\limsup_{n\to\infty}\frac{F\left(h+t_nH_n,c\right)-F\left(h,c\right)}{t_n}&\leqslant\lim_{n\to\infty}\frac{F\left(h+t_n\left(H+\zeta\right),c\right)-F\left(h,c\right)}{t_n}\\&=\int_{h^{-1}(c)}\frac{\left(H\left(x\right)+\zeta\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x.\end{aligned}$$

Similarly,

$$\begin{aligned}\liminf_{n\to\infty}\frac{F\left(h+t_nH_n,c\right)-F\left(h,c\right)}{t_n}&\geqslant\lim_{n\to\infty}\frac{F\left(h+t_n\left(H-\zeta\right),c\right)-F\left(h,c\right)}{t_n}\\&=\int_{h^{-1}(c)}\frac{\left(H\left(x\right)-\zeta\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x.\end{aligned}$$

By the Hölder inequality (integration by Hausdorff measure is also in the Lebesgue sense):

$$\int_{h^{-1}(c)}\frac{\left|H\left(x\right)-H'\left(x\right)\right|f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x\leqslant\sup_{x\in E}\left|H\left(x\right)-H'\left(x\right)\right|\int_{h^{-1}(c)}\left|\frac{f\left(x\right)}{Jh\left(x\right)}\right|d\mathcal{H}_{n-1}x.$$

Now by the arbitrariness of ζ,

$$\lim_{n\to\infty}\frac{F\left(h+t_nH_n,c\right)-F\left(h,c\right)}{t_n}=\int_{h^{-1}(c)}\frac{H\left(x\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x.$$

step 5 Now we can also prove Proposition 3.2.3. Note that the above proof techniques of **step 1** to **step** 3, including the area formula, partition of unity, and generalized matrix determinant lemma, all work for $k>1$. As a result, the proof is essentially a reconstruction of the proof of Theorem 3.2.2. Here, we only need to point out that

$$\begin{aligned}&\lim_{c'\to c}\int_{B_{x_{j,2},\ldots,x_{j,n}}}\frac{f\left(\psi_{B_j}\left(x_{2,\ldots,n},c'\right)\right)}{\left|\det\left(\frac{\partial}{\partial x_1}h\left(\psi_{B_j}\left(x_{2,\ldots,n},c'\right)\right)\right)\right|}d\mathcal{L}_{n-1}x\\&\quad=\int_{B_{x_{j,2},\ldots,x_{j,n}}}\frac{f\left(\psi_{B_j}\left(x_{2,\ldots,n},c\right)\right)}{\left|\det\left(\frac{\partial}{\partial x_1}h\left(\psi_{B_j}\left(x_{2,\ldots,n},c\right)\right)\right)\right|}d\mathcal{L}_{n-1}x.\end{aligned}$$

part 2 For $k>1$, we still have, by the coarea formula

$$F\left(h,c\right)=\int_{h(x)>c}f\left(x\right)d\mathcal{L}_nx=\int_{c'>c}\left[\int_{h^{-1}(c')}\frac{f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-k}\right]d\mathcal{L}_kc'.$$

So for all n large enough

$$\begin{aligned}&\frac{1}{t_n}\int\left[1\left(h\left(x\right)+t_nH_n\left(x\right)>c\right)-1\left(h\left(x\right)>c\right)\right]f\left(x\right)d\mathcal{L}_nx\\&=\frac{1}{t_n}\int\left[\int_{h^{-1}(c')\cap K_f}\frac{\left[1\left(c'+t_nH_n\left(x\right)>c\right)-1\left(c'>c\right)\right]f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-k}x\right]d\mathcal{L}_kc'.\end{aligned}$$

By the telescoping identity of higher order expansion,

$$\prod_{i=1}^{k}a_i'-\prod_{i=1}^{k}a_i=\sum_{i=1}^{k}\left(a_i'-a_i\right)\prod_{l\neq i}a_l+\sum_{i\neq j}\left(a_i'-a_i\right)\left(a_j'-a_j\right)\prod_{l\neq i,l\neq j}a_l+\cdots+\prod_{i=1}^{k}\left(a_i'-a_i\right),$$

using a proof procedure similar to that of Theorem 3.2.2, the first order term of the difference becomes

$$\begin{aligned}&\sum_i\frac{1}{t_n}\int\left[\int_{h^{-1}(c')\cap K_f}\frac{\left[1\left(c_i'+t_nH_{n,i}\left(x\right)>c_i\right)-1\left(c_i'>c_i\right)\right]\prod_{l\neq i}1\left(c_l'>c_l\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-k}x\right]d\mathcal{L}_kc'\\&=\sum_i\sum_{l_i}\int_{y>\tau_{\neg i}(c)}\int_{B_{\neg l_i}}d\mathcal{L}_{n-k}xdL_{k-1}y\\&\left[\int\frac{\left[1\left(c_i'+t_nH_{n,i}\left(\psi_{B_{l_i}}\left(x,c'\right)\right)>c_i\right)-1\left(c_i'>c_i\right)\right]f\left(\psi_{B_{l_i}}\left(x,c'\right)\right)\rho\left(\psi_{B_{l_i}}\left(x,c'\right)\right)}{\left|\det\left(\frac{\partial}{\partial x_{l_i}}h\left(\psi_{B_{l_i}}\left(x,c'\right)\right)\right)\right|}\frac{1}{t_n}d\mathcal{L}_1c_i'\right].\end{aligned}$$

where $\rho\left(\psi_{B_{l_i}}\left(x,c'\right)\right)$ is the partition of unity, and x_{l_i} denotes that x_{l_i} is picked for locally explicit function for change of variables by the implicit function theorem. We also note that in the above $y=\tau_{\neg i}\left(c'\right)$, and $c_i'=\tau_i\left(c'\right)$ and we use the fact that

$$1\left(c'+t_nH_n\left(x\right)>c\right)=\prod_{i=1}^{k}1\left(c_i'+t_nH_{n,i}\left(x\right)>c_i\right)\text{ and }1\left(c'>c\right)=\prod_{i=1}^{k}1\left(c_i'>c_i\right).$$

Then by DCT and the area formula again, we obtain the limit of the first order difference as

$$\sum_i\int\left[\int_{h^{-1}(c'[i])}\frac{H_i\left(x\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-k}x\right]d\mathcal{L}_{k-1}y.$$

The convergence of the first order term also implies that the second and higher order terms of the difference should vanish eventually.

part 3 By **part** 1 and **part** 2, we only need to consider the $k=1$ case since the $k>1$ case is similar. Let $\{c_n\}_{n=1}^{\infty}$ be a sequence such that $\lim_{n\to\infty} c_n = c$. Consider

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{t_n}\left[F\left(h+t_nH_n,c_n\right)-F\left(h,c_n\right)\right]\\&\quad=\lim_{n\to\infty}\frac{1}{t_n}\int\left[1\left(h\left(x\right)+t_nH_n\left(x\right)>c_n\right)-1\left(h\left(x\right)>c_n\right)\right]f\left(x\right)d\mathcal{L}_nx.\end{aligned}$$

The proof in **part** 1 shows that we need to calculate, for $M=\sup_{x\in K_f}|H\left(x\right)|$,

$$\frac{1}{t_n}\int_{c-t_nM-|c-c_n|}^{c+t_nM+|c-c_n|}\left[1\left(c'+t_nH\left(\psi_{B_j}\left(\cdot,c'\right)\right)>c_n\right)-1\left(c'>c_n\right)\right]d\mathcal{L}_1c'.$$

Note that

$$\frac{1}{t_n}\int\left[1\left(c'+t_nA+c-c_n>c\right)-1\left(c'+c-c_n>c\right)\right]d\mathcal{L}_1c'=A.$$

we can obtain

$$\lim_{n\to\infty}\frac{1}{t_n}\int_{c-t_nM-|c-c_n|}^{c+t_nM+|c-c_n|}\left[1\left(c'+t_nH\left(\psi_{B_j}\left(\cdot,c'\right)\right)>c_n\right)-1\left(c'>c_n\right)\right]d\mathcal{L}_1c'=H\left(\psi_{B_j}\left(\cdot,c\right)\right)$$

by simple upper and lower bound of $H\left(\psi_{B_j}\left(\cdot,c'\right)\right)$ on $[c-t_nM-|c-c_n|,c+t_nM+|c-c_n|]$. So we obtain, following the idea in Chernozhukov et al. (2018b), that the limit in the calculation of the Hadamard derivative actually converges continuously. By Lemma T.1.12, we get that the Hadamard derivative can be taken in a uniform sense. In other words, let $\mathcal{D}\subset h\left(E\right)$, then we have

$$\lim_{n\to\infty}\sup_{c\in\mathcal{D}}\left|\frac{1}{t_n}\int\left[1\left(h\left(x\right)+t_nH_n\left(x\right)>c\right)-1\left(h\left(x\right)>c\right)\right]f\left(x\right)d\mathcal{L}_nx-\int_{h^{-1}(c)}\frac{H\left(x\right)f\left(x\right)}{Jh\left(x\right)}d\mathcal{H}_{n-1}x\right|=0.$$

This finishes the proof of Theorem 3.2.1. ■

3.3 Binary treatment allocation and the ROC curve

In this subsection, we derive asymptotic pointwise and uniform distributions for binary treatment allocation value function under a resource constraint and for the estimated ROC curve. Here, using Hadamard differentiability results from subsection 3.2, we can directly analyze the plug-in two-step estimator of the value function for the constraint optimal allocation problems. Our strategy is concise. The functional delta methods alone (see for example van der Vaart and Wellner (2023) and Fang and Santos (2019)) are sufficient to obtain asymptotic results. The discussions in this subsection are conducted under the classical empirical process theory in the Hoffman-Jørgensen-Dudley sense. A most applicable result in this subsection is a computationally feasible delta method for bootstrap in probability consistency of the plug-in two-step ROC estimator which is commonly used in computer science and other fields.

Alternative theories can also be developed under other settings from different perspectives. In subsection 4.3, we visit the recent double/debiased machine learning methodology for unconstrained multi-classification problems. Double/debiased method under resource constraints can also be developed. We also note that the results in this subsection and in subsection 4.3 can be further improved to demonstrate semiparametric efficiency, which is analyzed in a follow-up manuscript Feng et al. (2024).

In binary applications of optimal constrained treatment allocation, we consider

$$\begin{aligned} &\max_{\phi\in\Phi}\mathbb{E}\left[Y_1\phi\left(X\right)+Y_0\left(1-\phi\left(X\right)\right)\right],\\ &s.t.\ \mathbb{E}\left[Z_1\phi\left(X\right)+Z_0\left(1-\phi\left(X\right)\right)\right]\leqslant\alpha,\end{aligned}$$

where Φ is the set of critical functions when $J=1$. Obviously, the problem above is equivalent to

$$\begin{aligned} &\max_{\phi\in\Phi}\mathbb{E}\left[\left(Y_1-Y_0\right)\phi\left(X\right)+Y_0\right],\\ &s.t.\ \mathbb{E}\left[\left(Z_1-Z_0\right)\phi\left(X\right)+Z_0\right]\leqslant\alpha.\end{aligned}\tag{13}$$

We will use the definition in (13) hereafter. By the classical Neyman-Pearson lemma (see, for example Theorem 3.2.1 in Lehmann and Romano (2022)), for α allowing feasible ϕ, the solution of the population optimization problem (13) exists and satisfies

$$\phi\left(x\right)=\begin{cases}1, & \text{when }\ \left(g_1\left(x\right)-g_0\left(x\right)\right)>k\left(c_1\left(x\right)-c_0\left(x\right)\right),\\ \\ 0, & \text{when }\ \left(g_1\left(x\right)-g_0\left(x\right)\right)<k\left(c_1\left(x\right)-c_0\left(x\right)\right),\end{cases}$$

for some constant k, where $g_1\left(x\right)=\mathbb{E}\left(Y_1|X=x\right)$, $g_0\left(x\right)=\mathbb{E}\left(Y_0|X=x\right)$, $c_1\left(x\right)=\mathbb{E}\left(Z_1|X=x\right)$ and $c_0\left(x\right)=\mathbb{E}\left(Z_0|X=x\right)$. When the fiber $\{x:\left(g_1\left(x\right)-g_0\left(x\right)\right)=k\left(c_1\left(x\right)-c_0\left(x\right)\right)\}$ is of zero probability mass, we can express the constraint and the value function of (13) as

$$\begin{aligned}\alpha\left(k\right)&=Q\left[\left(z_1-z_0\right)1\left(\left(g_1\left(x\right)-g_0\left(x\right)\right)>k\left(c_1\left(x\right)-c_0\left(x\right)\right)\right)+z_0\right]\leqslant\alpha,\\ \beta\left(k\right)&=Q\left[\left(y_1-y_0\right)1\left(\left(g_1\left(x\right)-g_0\left(x\right)\right)>k\left(c_1\left(x\right)-c_0\left(x\right)\right)\right)+y_0\right].\end{aligned}\tag{14}$$

Here, we adopt a functional notation where $\left(Y_1,Y_0,Z_1,Z_0,X\right)\sim Q$ and $X\sim\mu$. The Hadamard differentiability results established in subsection 3.2 allow us to derive asymptotic distribution of the sorting operator expression in (14) by the functional delta method. Denote $\ell^{\infty}\left(E\right)$ as the set of bounded real-valued functions on a set E. We need the following assumptions to validate a functional delta method. We largely adopt the styles in Chernozhukov et al. (2018b).

Assumption 3.3.1. *The random variables* $(Y_1, Y_0, Z_1, Z_0, X) \sim Q$, $X \sim \mu$, *where* μ *is absolutely continuous with respect to the* $\dim(X)$*-dimensional Lebesgue measure. The density* μ' *is continuous on an open set* $E \subset \mathbb{R}^{\dim(X)}$ *and* $\operatorname{supp}\mu' = K_{\mu'} \subset E$. (Y_1, Y_0, Z_1, Z_0) *is bounded and* $Z_1 \geqslant Z_0$.

Assumption 3.3.2. *The estimators* $(\hat{g}_0(x), \hat{g}_1(x), \hat{c}_0(x), \hat{c}_1(x))^T$ *satisfy*

$$r_n \left((\hat{g}_0(x), \hat{g}_1(x), \hat{c}_0(x), \hat{c}_1(x))^T - (g_0^\star(x), g_1^\star(x), c_0^\star(x), c_1^\star(x))^T \right) \rightsquigarrow \mathbb{H} \text{ in } \ell(E)^4,$$

where $r_n \to \infty$ *as* $n \to \infty$, $\mathbb{H}$ *is separable and its support is included in* $C(E)^4$.

Assumption 3.3.3. *The population limits satisfy* $(g_0^\star(\cdot), g_1^\star(\cdot), c_0^\star(\cdot), c_1^\star(\cdot)) \in C^1(E)^4$, $c_1^\star(x) > c_0^\star(x)$, *and* 0 *is a regular value of*

$$\Delta^\star(x;k) := (g_1^\star(x) - g_0^\star(x)) - k(c_1^\star(x) - c_0^\star(x)),$$

for all $k \in \mathcal{D}$, $\mathcal{D} \subset \mathbb{R}$ *a bounded closed interval.*

Assumption 3.3.4. *Let* $(g_0^\star(x), g_1^\star(x), c_0^\star(x), c_1^\star(x))^T \in \mathcal{G}$ *and* $(\hat{g}_0(x), \hat{g}_1(x), \hat{c}_0(x), \hat{c}_1(x))^T \in \mathcal{G}$ *with outer probability tends to* 1, *where* $\mathcal{G} \subset C(E)^4$. *The function class*

$$\mathcal{F} := \left\{ 1\left(\left(g_1'(x) - g_0'(x) \right) > k\left(c_1'(x) - c_0'(x) \right) \right) : \left(g_0'(\cdot), g_1'(\cdot), c_0'(\cdot), c_1'(\cdot) \right) \in \mathcal{G}, k \in \mathcal{D} \right\}$$

is Donsker, and the estimators and the empirical process converge jointly, .i.e.

$$t_n \left[\begin{pmatrix} \hat{g}_0(x) \\ \hat{g}_1(x) \\ \hat{c}_0(x) \\ \hat{c}_1(x) \end{pmatrix} - \begin{pmatrix} g_0^\star(x) \\ g_1^\star(x) \\ c_0^\star(x) \\ c_1^\star(x) \end{pmatrix}, \mathbb{Q}_n - Q \right] \rightsquigarrow (t_\Delta \mathbb{H}, t_Q \mathbb{Q}) \;\text{ in } C(E)^4 \times \ell^\infty(\mathcal{F}).$$

In the above, $t_n = \min\{r_n, \sqrt{n}\}$ *and* $\lim_{n\to\infty} \frac{t_n}{r_n} = t_\Delta \in [0,1]$, $\lim_{n\to\infty} \frac{t_n}{\sqrt{n}} = t_Q \in [0,1]$.

Theorem 3.3.1. *Let Assumptions 3.3.1 to 3.3.4 hold. Assume that there exists a constant* $\epsilon > 0$, *such that*

$$\int_{\Delta^\star(x;k)=0} \frac{(c_1(x) - c_0(x))(c_1^\star(x) - c_0^\star(x))\mu'(x)}{\|\nabla\Delta^\star(x;k)\|} d\mathcal{H}_{n-1}x > \epsilon$$

for all $k \in \mathcal{D}$*. Then, as a stochastic process on compact* $\Lambda' \subsetneq \mathring{\Lambda}$, $\Lambda := Q\left[(z_1 - z_0) 1(\Delta(x;\mathcal{D}) > 0)\right]$,

$$
\begin{aligned}
& t_n\left(\beta\left(\mathbb{Q}_n, \hat{\Delta}, \alpha\right) - \beta\left(Q, \Delta^\star, \alpha\right)\right) \rightsquigarrow \Bigg\{ t_Q \mathbb{Q}\left[(y_1 - y_0) 1\left(\Delta^\star\left(x; k\left(Q, \Delta^\star, \alpha\right)\right) > 0\right) + y_0\right] \\
& + t_\Delta \Bigg[\int_{\Delta^\star(x;k(Q,\Delta^\star,\alpha))=0} \frac{\left(g_1(x) - g_0(x)\right)\left(\mathbb{H}_1 - \mathbb{H}_0\right)\mu'(x)}{\|\nabla\Delta^\star\left(x; k\left(Q, \Delta^\star, \alpha\right)\right)\|} d\mathcal{H}_{n-1}x \\
& \qquad - \int_{\Delta^\star(x;k(Q,\Delta^\star,\alpha))=0} \frac{\left(g_1(x) - g_0(x)\right)\left(\mathbb{H}_3 - \mathbb{H}_2\right)\mu'(x)}{\|\nabla\Delta^\star\left(x; k\left(Q, \Delta^\star, \alpha\right)\right)\|} d\mathcal{H}_{n-1}x \Bigg] \Bigg\} \\
& - \frac{f_\beta\left(Q, \Delta^\star, k\left(Q, \Delta^\star, \alpha\right)\right)}{f_\alpha\left(Q, \Delta^\star, k\left(Q, \Delta^\star, \alpha\right)\right)} \Bigg\{ t_Q \mathbb{Q}\left[(z_1 - z_0) 1\left(\Delta^\star\left(x; k\left(Q, \Delta^\star, \alpha\right) > 0\right)\right) + z_0\right] \\
& + t_\Delta \Bigg[\int_{\Delta^\star(x;k(Q,\Delta^\star,\alpha))=0} \frac{\left(c_1(x) - c_0(x)\right)\left(\mathbb{H}_1 - \mathbb{H}_0\right)\mu'(x)}{\|\nabla\Delta^\star\left(x; k\left(Q, \Delta^\star, \alpha\right)\right)\|} d\mathcal{H}_{n-1}x \\
& \qquad - \int_{\Delta^\star(x;k(Q,\Delta^\star,\alpha))=0} \frac{\left(c_1(x) - c_0(x)\right)\left(\mathbb{H}_3 - \mathbb{H}_2\right)\mu'(x)}{\|\nabla\Delta^\star\left(x; k\left(Q, \Delta^\star, \alpha\right)\right)\|} d\mathcal{H}_{n-1}x \Bigg] \Bigg\},
\end{aligned} \tag{15}
$$

where $\left(\mathbb{H}_0, \mathbb{H}_1, \mathbb{H}_2, \mathbb{H}_3\right)^T = \mathbb{H}$*;*

$$
\begin{aligned}
f_\beta\left(Q, \Delta^\star, k\right) &:= \frac{d\beta\left(Q, \Delta^\star, k\right)}{dk} = -\int_{\Delta^\star(x;k)=0} \frac{\left(g_1(x) - g_0(x)\right)\left(c_1^\star(x) - c_0^\star(x)\right)\mu'(x)}{\|\nabla\Delta^\star(x;k)\|} d\mathcal{H}_{n-1}x, \\
f_\alpha\left(Q, \Delta^\star, k\right) &:= \frac{d\alpha\left(Q, \Delta^\star, k\right)}{dk} = -\int_{\Delta^\star(x;k)=0} \frac{\left(c_1(x) - c_0(x)\right)\left(c_1^\star(x) - c_0^\star(x)\right)\mu'(x)}{\|\nabla\Delta^\star(x;k)\|} d\mathcal{H}_{n-1}x.
\end{aligned}
$$

The notation Δ $(\hat{\Delta}, \Delta^\star)$ *denotes* $\left(g_0(\cdot), g_1(\cdot), c_0(\cdot), c_1(\cdot)\right)$ *(their estimators and the estimators' limits), and*

$$
\begin{aligned}
& \beta(Q, \Delta, \alpha) := \beta(Q, \Delta, k(Q, \Delta, \alpha)) = Q\left[(y_1 - y_0) 1(\Delta(x; k(Q, \Delta, \alpha)) > 0) + y_0\right], \\
& \alpha(Q, \Delta, k) := Q\left[(z_1 - z_0) 1(\Delta(x;k) > 0) + z_0\right],
\end{aligned}
$$

$k(Q, \Delta, \alpha) = \inf\{k \in \mathbb{R} : \alpha(Q, \Delta, k) \leqslant \alpha\}$.

Proof of Theorem 3.3.1. Note that under Assumption 3.3.1 and Assumption 3.3.3, we can evoke Theorem 3.2.1; under Assumption 3.3.4 we can evoke Lemma A.2 in Chernozhukov et al. (2018b). Also by the chain rule of Hadamard derivative (Lemma 3.10.3 in van der Vaart and

Wellner (2023)),

$$\alpha'_{Q,\Delta^\star,k}(\mathcal{Q},H,0) = \mathcal{Q}\left[(z_1-z_0)\,1\,(\Delta^\star(x;k)>0)+z_0\right]$$
$$+\left[\int_{\Delta^\star(x;k)=0}\frac{(c_1(x)-c_0(x))\,(H_1(x)-H_0(x))\,\mu'(x)}{\|\nabla\Delta^\star(x,k)\|}d\mathcal{H}_{n-1}x\right.$$
$$\left.-\int_{\Delta^\star(x;k)=0}\frac{(c_1(x)-c_0(x))\,(H_3(x)-H_2(x))\,\mu'(x)}{\|\nabla\Delta^\star(x,k)\|}d\mathcal{H}_{n-1}x\right].$$

By the assumption that

$$\int_{\Delta^\star(x;k)=0}\frac{(c_1(x)-c_0(x))\,(c_1^\star(x)-c_0^\star(x))\,\mu'(x)}{\|\nabla\Delta^\star(x;k)\|}d\mathcal{H}_{n-1}x$$

is uniformly bounded away from zero on $k\in\mathcal{D}$, the Hadamard differentiability of the inverse map in Lemma 3.10.24 of van der Vaart and Wellner (2023) and the chain rule again imply that

$$k'_{Q,\Delta^\star,\alpha}(\mathcal{Q},H,0)=\frac{\alpha'_{Q,\Delta^\star,k}(\mathcal{Q},H,0)}{f_\alpha(Q,\Delta^\star,k(Q,\Delta^\star,\alpha))},$$

where the Hadamard differentiability is tangential to $\mathbf{Q}\times C(E)$, such that $\mathbf{Q}\subset\ell^\infty(\mathcal{F})$ consists of elements that are uniformly continuous on $\mathcal{F}$ with respect to the $L^2(Q)$ norm.

By exactly the same calculation, we can also get

$$\beta'_{Q,\Delta^\star,k}(\mathcal{Q},H,0) = \mathcal{Q}\left[(y_1-y_0)\,1\,(\Delta^\star(x;k)>0)+y_0\right]$$
$$+\left[\int_{\Delta^\star(x;k)=0}\frac{(g_1(x)-g_0(x))\,(H_1(x)-H_0(x))\,\mu'(x)}{\|\nabla\Delta^\star(x,k)\|}d\mathcal{H}_{n-1}x\right.$$
$$\left.-\int_{\Delta^\star(x;k)=0}\frac{(g_1(x)-g_0(x))\,(H_3(x)-H_2(x))\,\mu'(x)}{\|\nabla\Delta^\star(x,k)\|}d\mathcal{H}_{n-1}x\right].$$

By the chain rule the third time, there is

$$\beta'_{Q,\Delta^\star,\alpha}(\mathcal{Q},H,0)=\beta'_{Q,\Delta^\star,k}(\mathcal{Q},H,0)\Big|_{k=k(Q,\Delta^\star,\alpha)}-\frac{f_\beta(Q,\Delta^\star,k(Q,\Delta^\star,\alpha))}{f_\alpha(Q,\Delta^\star,k(Q,\Delta^\star,\alpha))}\alpha'_{Q,\Delta^\star,k}(\mathcal{Q},H,0)\Big|_{k=k(Q,\Delta^\star,\alpha)}.$$

Now, (15) follows from Theorem 3.10.4 in van der Vaart and Wellner (2023). ∎

A useful special case of (13) is when $Y_1\in\{0,1\}$, $Y_0=Z_0=0$, $Z_1=1-Y_1$, and $g_0=c_0\equiv 0$, $g_1(x)=p(x)$, $c_1(x)=1-p(x)$ where $p(x)=\mathbb{E}[Y|\,X=x]$ with $Y=Y_1$. This special case, i.e.

$$\max_{\phi\in\Phi}\mathbb{E}[Y\phi(X)],\ s.t.\ \mathbb{E}[(1-Y)\phi(X)]\leqslant\alpha.$$

is closely related to the receiver operating characteristic (ROC) curve whose population definition is given by

$$\beta(\alpha) := \max_{\phi\in\Phi} \frac{\mathbb{E}[Y\phi(X)]}{\mathbb{E}Y}, \ s.t. \ \frac{\mathbb{E}[(1-Y)\phi(X)]}{\mathbb{E}(1-Y)} \leqslant \alpha.$$

In the following presentation of the uniform asymptotic distribution of the ROC curve, we differentiate between misspecified models and correctly specified models. The limit of misspecified models is essentially a direct application of the general results in Theorem 3.3.1. When the parametric propensity score model is correctly specified, or when it is estimated by nonparametric techniques with a sufficiently fast rate of convergence, it will be shown that the first stage estimation of the propensity score model has no impact on the asymptotic distribution of the ROC curve. This property simplifies the inference procedure when the propensity score is nonparametrically estimated or correctly parametrically specified. In particular, it is not necessary to reestimate the propensity score model in a bootstrap procedure.

Corollary 3.3.2. *Under the conditions in Theorem 3.3.1, but let* $(g_0(x), g_1(x), c_0(x), c_1(x)) = (0, p(x), 0, 1-p(x))$ *(with the corresponding expression in Assumption 3.3.1 to 3.3.4 adapting to the change of notations). Then we have*

$$\begin{aligned}
&t_n(\beta(\mathbb{Q}_n,\hat{p},\alpha)-\beta(Q,p^\star,\alpha)) \rightsquigarrow \frac{1}{Qy}\Bigg\{t_Q\mathbb{Q}[y(1(p^\star(x)>k(Q,p^\star,\alpha))-\beta(Q,p^\star,\alpha))]\\
&+t_\Delta\int_{p^{\star-1}(k(Q,p^\star,\alpha))}\frac{p(x)\mathbb{H}(x)\mu'(x)}{\|\nabla p^\star(x)\|}d\mathcal{H}_{n-1}x\Bigg\}-\frac{1}{Qy}\frac{f_\beta(Q,p^\star,k(Q,p^\star,\alpha))}{f_\alpha(Q,p^\star,k(Q,p^\star,\alpha))}\Bigg\{\\
&t_Q\mathbb{Q}[(1-y)(1(p^\star(x)>k(Q,p^\star,\alpha))-\alpha)]+t_\Delta\int_{p^{\star-1}(k(Q,p^\star,\alpha))}\frac{p(x)\mathbb{H}(x)\mu'(x)}{\|\nabla p^\star(x)\|}d\mathcal{H}_{n-1}x\Bigg\}
\end{aligned}$$

on compact $\Lambda' \subsetneq \mathring{\Lambda}$, $\Lambda := \frac{Q[(1-y)1(p^\star(x)>\mathcal{D})]}{Q(1-y)}$*. Specially, when* $p^\star$ *is correctly, i.e.* $p^\star(x) = p(x)$*, then we get*

$$\begin{aligned}
t_n(\beta(\mathbb{Q}_n,\hat{p},\alpha)-\beta(Q,p^\star,\alpha)) \rightsquigarrow \frac{t_Q}{Qy}\Bigg\{&\mathbb{Q}[y(1(p(x)>k(Q,p,\alpha))-\beta(Q,p,\alpha))]\\
&-\frac{k(Q,p,\alpha)}{1-k(Q,p,\alpha)}\mathbb{Q}[(1-y)(1(p(x)>k(Q,p,\alpha))-\alpha)]\Bigg\}.
\end{aligned}$$

Proof of Corollary 3.3.2. The whole proof is similar to the proof of Theorem 3.3.1, we only need to point out that by chain rule

$$\begin{aligned}
&\alpha'_{Q,p^\star,k}(\mathcal{Q},H,0)\\
&=\frac{1}{Q(1-y)}\left[\mathcal{Q}[(1-y)(1(p^\star(x)>k)-\alpha(Q,p^\star,k))]+\int_{p^{\star-1}(k)}\frac{(1-p(x))H(x)\mu'(x)}{\|\nabla p^\star(x)\|}d\mathcal{H}_{n-1}x\right],
\end{aligned}$$

so we get an additional α term. ■

Under correct model specification, corollary 3.3.2 implies a computationally feasible functional delta method for bootstrap in probability. Here, we do not need to recompute / reestimate the first step model for every bootstrapped sample. On the contrary, the first step only need to be trained once.

Corollary 3.3.3. *Under the conditions in Theorem 3.3.1, let*

$$\left(\hat{p}, \tilde{\mathbb{Q}}_n\right) := \left(\hat{p}\left(\{(X_i, Y_i)\}_{i=1}^n\right), \tilde{\mathbb{Q}}_n\left(\{(X_i, Y_i)\}_{i=1}^n, \{M_i\}_{i=1}^n\right)\right)$$

be the first step estimator and the empirical bootstrapped sample conditional on $\{(X_i, Y_i)\}_{i=1}^n$. *Assume also* $t_n = \sqrt{n}$, *then*

$$\sup_{l \in BL_1(\ell^\infty(\mathcal{D}))} \left| \mathbb{E}_M l\left(\sqrt{n}\left(\beta\left(\tilde{\mathbb{Q}}_n, \hat{p}, \alpha\right) - \beta\left(\mathbb{Q}_n, \hat{p}, \alpha\right)\right)\right) - \mathbb{E} l\left(\beta'_{Q,p,\alpha}\left(\mathbb{Q}, 0, 0\right)\right)\right| \to 0,$$

$$\mathbb{E}_M l\left(\sqrt{n}\left(\beta\left(\tilde{\mathbb{Q}}_n, \hat{p}, \alpha\right) - \beta\left(\mathbb{Q}_n, \hat{p}, \alpha\right)\right)\right)^* - \mathbb{E}_M l\left(\sqrt{n}\left(\beta\left(\tilde{\mathbb{Q}}_n, \hat{p}, \alpha\right) - \beta\left(\mathbb{Q}_n, \hat{p}, \alpha\right)\right)\right)_* \to 0$$

in outer probability.

Proof of Corollary 3.3.3. First note that by Assumption 3.3.4, we can evoke Theorem 3.7.1 in van der Vaart and Wellner (2023) to get

$$\sup_{l \in BL_1(\ell^\infty(E) \times \ell^\infty(\mathcal{F}))} \left| \mathbb{E}_M l\left(\sqrt{n}\left(\begin{pmatrix} \hat{p} \\ \tilde{\mathbb{Q}}_n \end{pmatrix} - \begin{pmatrix} \hat{p} \\ \mathbb{Q}_n \end{pmatrix}\right)\right) - \mathbb{E} l\left(\begin{pmatrix} 0 \\ \mathbb{Q} \end{pmatrix}\right)\right| \to 0,$$

$$\mathbb{E}_M l\left(\sqrt{n}\left(\begin{pmatrix} \hat{p} \\ \tilde{\mathbb{Q}}_n \end{pmatrix} - \begin{pmatrix} \hat{p} \\ \mathbb{Q}_n \end{pmatrix}\right)\right)^* - \mathbb{E}_M l\left(\sqrt{n}\left(\begin{pmatrix} \hat{p} \\ \tilde{\mathbb{Q}}_n \end{pmatrix} - \begin{pmatrix} \hat{p} \\ \mathbb{Q}_n \end{pmatrix}\right)\right)_* \to 0$$

in outer probability. The results then follow from Theorems 3.10.4 and 3.10.11 in van der Vaart and Wellner (2023). ■

4 Sub (sup) and Fréchet differentiability

We first recall the definition of sub (sup) differentiability. Let E be an open set of $\mathbb{R}^n$, and $f : E \to \mathbb{R}$ a function. Then f is said to be subdifferentiable at x, with subgradient p, if

$$f\left(x'\right) \geqslant f\left(x\right) + \left\langle p, x' - x \right\rangle + o\left(\left\|x' - x\right\|\right).$$

The convex set of all subgradients p at x will be denoted by $\nabla^{-} f\left(x\right)$. If $(-f)$ is subdifferentiable, then f is said to be supdifferentiable, and the convex set of the negated subgradients for $(-f)$ at x is denoted as $\nabla^{+} f$.[1] It is well known that a convex function f is subdifferentiable with

$$f\left(x'\right) \geqslant f\left(x\right) + \left\langle p, x' - x \right\rangle$$

for some convex set of p. A concave function $h\left(\cdot\right)$ is supdifferentiable with

$$h\left(x'\right) \leqslant h\left(x\right) + \left\langle p, x' - x \right\rangle$$

for some convex set of p.

4.1 Social welfare potential function

Given the convexity in λ given $g\left(\cdot\right)$, a subgradient of $\gamma\left(\lambda, g\right)$ in λ is given by

$$p\left(\lambda\right) = \left(\mathbb{E}\phi_0^*\left(X; \lambda, g\right) g_0\left(X\right), \ldots, \mathbb{E}\phi_J^*\left(X; \lambda, g\right) g_J\left(X\right)\right)^T. \tag{16}$$

where $\phi^*\left(\cdot; \lambda, g\right)$ is one of the optimal allocation under the parameter (λ, g). To verify by direct calcuation,

$$\begin{aligned}
&\gamma\left(\lambda, g\right) + \left\langle p\left(\lambda\right), \lambda' - \lambda \right\rangle \\
&= \sum_j \mathbb{E}\lambda_j \phi_j^*\left(X; \lambda, g\right) g_j\left(X\right) + \sum_j \mathbb{E}\left(\lambda_j' - \lambda_j\right) \phi_j^*\left(X; \lambda, g\right) g_j\left(X\right) \\
&= \sum_j \mathbb{E}\lambda_j' \phi_j^*\left(X; \lambda, g\right) g_j\left(X\right) \leqslant \sum_j \mathbb{E}\lambda_j' \phi_j^*\left(X; \lambda', g\right) g_j\left(X\right) = \gamma\left(\lambda', g\right).
\end{aligned}$$

The rest of this subsection rigorously justifies calling $\gamma\left(\lambda, g\right)$ a social welfare *potential* function, with $p\left(\lambda\right)$ belonging to the vector field generated by $\gamma\left(\lambda, g\right)$.

Recall by the concept of subgradients the set-valued map $\nabla\gamma\left(\lambda, g\right) : \mathbb{R}^{J+1} \rightrightarrows \mathbb{R}^{J+1}$, where the symbol $\rightrightarrows$ indicates that the image of a point under the map is a set. A common function is then understood as a set-valued map where the image of a single point is a singleton set. The following definition is needed:

[1] We follow the notation convention in Villani et al. (2009).

Definition 4.1.1. (Conservative set-valued fields, Bolte and Pauwels (2021)) Let $D : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a set- valued map. D is a conservative (set-valued) field whenever it has closed graph, nonempty compact values, and for any absolutely continuous loop $\ell : [0,1] \to \mathbb{R}^p$, such that $\ell(0) = \ell(1)$, the Aumann integral of $t \rightrightarrows \left\langle \dot{\ell}(t), D(\ell(t)) \right\rangle$ is $\{0\}$, namely,

$$\begin{aligned} &\int_0^1 \left\langle \dot{\ell}(t), D(\ell(t)) \right\rangle dt \\ &:= \left\{ \int_0^1 \omega(t) dt : \omega(t) : [0,1] \to R \text{ is a measurable selection of} \left\langle \dot{\ell}(t), D(\ell(t)) \right\rangle \right\} = \{0\}. \end{aligned} \tag{17}$$

It is shown in Bolte and Pauwels (2021) (17) is equivalent to requiring that

$$\int_0^1 \max_{v \in D(\ell(t))} \left\langle \dot{\ell}(t), v \right\rangle dt = 0,$$

where the integral is understood in the Lebesgue sense, which is possible by Theorem 18.19 and Theorem 18.20 in Aliprantis and Border (2006). See also Lemma 1 in Bolte and Pauwels (2021). Another equivalent requirement is

$$\int_0^1 \min_{v \in D(\ell(t))} \left\langle \dot{\ell}(t), v \right\rangle dt = 0 \quad \text{for all absolutely continuous loop } \ell.$$

The definition above involves an argmax (argmin) measurable selection, so for any measurable selection $v(t) : [0,1] \to \mathbb{R}^p, v(t) \in D(\ell(t))$, we have

$$\int_0^1 \left\langle \dot{\ell}(t), v(t) \right\rangle dt = 0.$$

Based on Definition (4.1.1), we introduce the following Definition (4.1.2):

Definition 4.1.2. (Potential functions of conservative fields) A function $f : \mathbb{R}^p \to \mathbb{R}$ is called a potential function of a conservative field whenever there exists a conservative field $D : \mathbb{R}^p \to \mathbb{R}^p$ such that

$$f(x) = f(0) + \int_0^1 \langle \dot{\mathtt{p}}(t), D(\mathtt{p}(t)) \rangle dt$$

for all absolutely continuous path $\mathtt{p} : [0,1] \to \mathbb{R}^p$ with $\mathtt{p}(0) = 0$ and $\mathtt{p}(1) = x$.

Bolte and Pauwels (2021) shows that a potential function must be locally Lipschitz and their Theorem 1 states that if a function f is a potential function of the conservative field D then D coincides with the gradient of f on Lebesgue almost every point where the gradient exists. Therefore, we can expect that certain kind of nonsmooth functions will generate conservative fields by some generalized approach of taking derivatives. In addition, a real-valued convex (or concave) function is locally Lipschitz and thus Lebesgue almost everywhere differentiable by the Rademacher's theorem. See for example Theorem 3.7.3 in Niculescu and Persson (2018) and also Theorem T.1.2 in the technical addendum T.

Theorem 4.1.1. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a convex (or concave) function, then $\nabla^- f\,(\nabla^+ f) = \partial f$ is a conservative field, where ∂f is the Clarke subgradient defined below.*

Definition 4.1.3. (Clarke subgradients) Consider a local Lipschitz function $f : \Omega \to R$, where $\Omega \subset \mathbb{R}^n$ is an open subset. For each $x \in \Omega$, define

$$\begin{aligned} f^\circ(x; v) &:= \limsup_{\substack{y \to x \\ \lambda \downarrow 0}} \frac{f(y + \lambda v) - f(y)}{\lambda} \\ &= \lim_{\substack{\varepsilon \to 0 \\ \epsilon \downarrow 0}} \sup \left\{ \frac{f(y + \lambda v) - f(y)}{\lambda} : y \in \Omega \cap B(x, \varepsilon), \lambda \in (0, \epsilon) \right\}, \forall v \in \mathbb{R}^n, y + \lambda v \in \Omega, \end{aligned}$$

and the Clarke subgradient of f at x:

$$\partial f(x) := \{\xi \in \mathbb{R}^n : f^\circ(x; v) \geqslant \langle v, \xi \rangle, \forall v \in \mathbb{R}^n\}\,[2].$$

Clarke (1975) provides an alternative characterization of Clarke subgradients in $\mathbb{R}^n$: Let $f : \Omega \to R$ be a function in Definition 4.1.3 , then

$$\partial f(x) = \operatorname{conv} \left\{ \lim_{k \to \infty} \nabla f(x_k) : x_k \to x, \nabla f(x_k) \text{ exists} \right\}.$$

Proof of Theorem 4.1.1. By Proposition 1.2 in Clarke (1975), for convex (concave) functions f, $\nabla^- f\,(\nabla^+ f) = \partial f$. Furthermore, $\nabla^- f(x)\,(\nabla^+ f(x)) = \partial_g f(x)$ by Proposition 8.12 in Rockafellar and Wets (2009), where $\partial_g f(x)$ is the general subgradient as in definition 8.3 in Rockafellar and Wets (2009). Finally, by Theorem 10.49 in Rockafellar and Wets (2009) and Corollary 2 in Bolte and Pauwels (2021), ∂f is a conservative field. ∎

4.2 Envelope like theorem for social welfare potential function

Consider, for a given $g(\cdot)$,

$$e\left(\lambda'; \phi^*(\cdot; \lambda, g)\right) = \mathbb{E}\left[\sum_j \lambda'_j \phi^*_j g_j\right] - \gamma\left(\lambda', g\right),$$

by construction, $e(\lambda'; \phi^*(\cdot; \lambda, g))$ achieves its maximum of zero at $\lambda' = \lambda$. When there is no arbitrariness, we also denote $\phi^*(\cdot) := \phi^*(\cdot; \lambda, g)$ Note first that $\mathbb{E}\left[\sum_j \lambda'_j \phi^*_j g_j\right]$ is everywhere differentiable in λ'_j given $\phi^*(\cdot)$ and $g(\cdot)$. Because of the convexity of $\gamma(\lambda', g)$ in λ' given $g(\cdot)$, $\gamma(\lambda', g)$ is differentiable in λ' given $g(\cdot)$ in a set Λ such that its complement Λ^c is a Lebesgue null set. Then for each $\lambda \in \Lambda$, and for each selection ϕ^* given this choice of λ, $e(\lambda'; \phi^*)$ is differentiable in λ' at $\lambda' = \lambda$. Since zero must be in the set of supgradients at a point of maximum of any function, there is $0 \in \nabla^+ e(\lambda'; \phi^*)\big|_{\lambda'=\lambda}$ and thus $\nabla e(\lambda'; \phi^*)\big|_{\lambda'=\lambda} = 0$, implying

$$\nabla\gamma(\lambda', g)\Bigg|_{\lambda'=\lambda} = \nabla\mathbb{E}\left[\sum_j \lambda'_j \phi^*_j g_j\right]\Bigg|_{\lambda'=\lambda} = \left(\mathbb{E}\phi^*_j g_j, j = 0, \ldots, J\right)^T.$$

An additional implication of the previous result is that for each $\lambda \in \Lambda$, $\phi^*(x; \lambda, g)$ is unique in an almost surly sense since the expectation $\Big(\mathbb{E}\phi^*_j g_j, j = 0, \ldots, J\Big)^T$ is uniquely defined.

Intuitively, we expect such envelope theorem like statement remains true with respect to certain functional derivatives, so that we can discuss the functional derivatives of

$$e(\lambda', g'; \phi^*) = \mathbb{E}\left[\sum_j \lambda'_j \phi^*_j g'_j\right] - \gamma(\lambda', g') \tag{18}$$

at $(\lambda', g') = (\lambda, g)$ (here, $\phi^*(\cdot) := \phi^*(\cdot; \lambda, g)$). This conjecture will be subsequently verified. Surprisingly, some Fréchet differentiability arguments can be established.

Definition 4.2.1. (Fréchet differentiability) Let $\mathcal{X}$ and $\mathcal{Y}$ be normed spaces equipped with norm $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$, $E \subset \mathcal{X}$ be an open set. Consider the map $\xi : E \to Y$. Then ξ is called Fréchet differentiable at $\theta \in E$, if there is a continuous linear map $\xi'_\theta : \mathcal{X} \to \mathcal{Y}$ such that:

$$\left\|\xi(x) - \xi(\theta) - \xi'_\theta(x - \theta)\right\|_{\mathcal{Y}} = o\left(\|x - \theta\|_{\mathcal{X}}\right).$$

Before we describe the next result we need to recall several topological and measure theoretic concepts. In a topological space $\mathcal{X}$, if a subset $S \in \mathcal{X}$ satisfies $\mathring{\bar{S}} = \varnothing$, i.e. the closure of S has empty interior, then S is called a nowhere dense subset of $\mathcal{X}$. If $\mathcal{Z} \subset \mathcal{X}$ is a countable union of nowhere dense subsets of $\mathcal{X}$, then $\mathcal{Z}$ is called a meager subset of X, or of the first category in $\mathcal{X}$. A subset $A \subset \mathcal{X}$ is called residually many (or just residual) in $\mathcal{X}$ if $\mathcal{X}\backslash A$ is meager in $\mathcal{X}$. A set $B \subset \mathcal{X}$ is called a G_δ set of $\mathcal{X}$ if B is a countable intersection of open sets; it is call a F_σ set if it is a countable union of closed sets.

In Lindenstrauss and Preiss (2003), the authors originate an elaborate concept to describe the *magnitude* of sets, call Γ-null sets. Let $T = [0,1]^{\mathbb{N}}$, where $\mathbb{N}$ is the set of natural numbers, be endowed with the product topology and product Lebesgue measure $\mathcal{L}_\infty$. By the Tikhonov

Theorem, T is a compact space. It can also be metrized by

$$\rho(x,y)=\sum_{n=1}^{\infty}\frac{\rho_n(x_n,y_n)}{2^n(1+\rho_n(x_n,y_n))}$$

where ρ_n is some distance on $[0,1]$.

By the product σ-algebra on the space $\prod_i E_i$, we mean the σ-algebra generated by all the coordinate projections. The product measure then should satisfy

$$\mu\left[\prod_i B_i, B_i\subset\mathcal{E}_i,\ \text{and } B_i\neq E_i, \forall i\in\mathcal{I}, B_i=E_i, \forall i\in\mathbb{N}\backslash\mathcal{I},\ \text{where } \mathcal{I}\subset\mathbb{N}, 0<|I|<\infty\right]=\prod_{i\in\mathcal{I}}\mu_i(B_i),$$

where $\{(E_i,\mathcal{E}_i,\mu_i)\}$ is a family of measure spaces. The existence and uniqueness of the countable product Lebesgue measure is guaranteed by the Kolmogorov extension theorem (See for example Chapter 1 and 3 in Dellacherie and Meyer (1978)).

Now, let $\mathcal{X}$ be a Banach space and $\Gamma(\mathcal{X})=\{\gamma:T\to\mathcal{X}\}$ be the space of continuous mappings having continuous partial derivatives $D_j\gamma$. Equip $\Gamma(\mathcal{X})$ with a topology generated by $||\gamma||_0=\sup_{t\in T}||\gamma(t)||$, and all $||\gamma||_k=\sup_{t\in T}||D_k\gamma(t)||$, $k\geqslant 1$. Equivalently, the same topology is generated by all $||\gamma||_{\leqslant k}=\max_{0\leqslant i\leqslant k}||\gamma||_i, k\in\mathbb{N}$.

Definition 4.2.2. (Γ-null) A Borel set $N\subset\mathcal{X}$ is called Γ-null if $\mathcal{L}_\infty\{t\in T:\gamma(t)\in N\}=0$ for residually many $\gamma\in\Gamma(\mathcal{X})$. If a set A is contained in such a N, then A is also called Γ-null.

Now, we can state our result. In the following theorem statement, 1 and 2 follow from the convexity of $\gamma(\lambda,g)$ in λ given each g, and the convexity of $\gamma(\lambda,g)$ in g given each each λ, while 3 is needed to account for the non joint-convexity of $\gamma(\lambda,g)$ in (λ,g).

Theorem 4.2.1. *Let $\gamma(\lambda,g)$ be a real valued social welfare potential function, where $\lambda\in\mathbb{R}^{J+1}$, $g\in\mathcal{X}$ where $\mathcal{X}$ is a Banach space. Assume that $\gamma(\lambda,g)$ is continuous at a point $(\lambda_0,g_0)\in\mathbb{R}^{J+1}\otimes\mathcal{X}$, then we have:*

1. *If every separable subspace $\mathcal{Y}$ of $\mathcal{X}$ has a separable dual space $\mathcal{Y}^*$, then for any given $\lambda\in\mathbb{R}^{J+1}$, $\gamma(\lambda,g)$ is Fréchet differentiable in g on a dense G_δ subset of $\mathcal{X}$.*

2. *If the dual space $\mathcal{X}^*$ of $\mathcal{X}$ is separable, then for any given $\lambda\in\mathbb{R}^{J+1}$, $\gamma(\lambda,g)$ is Γ-almost everywhere Fréchet differentiable with respect to $g\in\mathcal{X}$. In other words, for any given $\lambda\in\mathbb{R}^{J+1}$, $\gamma(\lambda,g)$ is not Fréchet differentiable in g on a Γ-null set of $\mathcal{X}$.*

3. *In addition, If every separable subspace $\mathcal{Y}$ of $\mathcal{X}$ has a separable dual space $\mathcal{Y}^*$, and $\mathcal{X}$ can be continuously embedded into $L^1(\Omega,\mu)$ (in the sense that for all $x\in\mathcal{X}$, $\|x\|_{L^1(\Omega,\mu)}\leqslant C\|x\|_{\mathcal{X}}$ for a constant C), then $\gamma(\lambda,g)$ is locally jointly Lipschitz in (λ,g), and is also jointly Fréchet*

differentiable with respect to $(\lambda, g(\cdot))$ *on a dense subset of* $\mathbb{R}^{J+1} \otimes \mathcal{X}$.

When $\gamma(\lambda, g)$ *is Fréchet differentiable (totally or partially as in the theorem), the Fréchet derivative with respect to* (λ, g) *can be calculated to be*

$$((\mathbb{E}\phi_0^* g_0, \ldots, \mathbb{E}\phi_J^* g_J), (\mathbb{E}\lambda_0 \phi_0^*, \ldots, \mathbb{E}\lambda_J \phi_J^*))^T \tag{19}$$

where

$$\phi^*(\cdot) := \phi^*(\cdot; \lambda, g) = \underset{\phi(\cdot)\in\Phi}{\arg\max}\, \mathbb{E}\left[\sum_j \lambda_j \phi_j(X) g_j(X)\right]$$

in the μ *almost surely sense, where* μ *is the distribution of* X.

PROOF OF THEOREM 4.2.1.

step 1 First, it is direct that $\gamma(\lambda, g)$ satisfies subadditivity and is also positive homogeneous of degree one with respect to both λ and $g(\cdot)$, separately,

$$\gamma(\lambda, g + g') \leqslant \gamma(\lambda, g) + \gamma(\lambda, g'), \quad \gamma(\lambda + \lambda', g) \leqslant \gamma(\lambda, g) + \gamma(\lambda', g),$$
$$\gamma(\alpha\lambda, g) = \alpha\gamma(\lambda, g), \quad \gamma(\lambda, \alpha g) = \alpha\gamma(\lambda, g), \text{ for all } \alpha \geqslant 0,$$

implying that $\gamma(\lambda, g)$ is convex in $g(\cdot)$ given λ and is convex in λ given $g(\cdot)$. Note that when $\mathcal{X}$ is infinite-dimensional, a real-valued convex function is not necessarily continuous (there is always a noncontinuous linear functional on $\mathcal{X}$). Fortunately, a real-valued convex function on a Banach space is either continuous at evey point or discontinuous at every point of its domain (see for example Proposition 3.1.11 of Niculescu and Persson (2018)). So by assumption, $\gamma(\lambda, g)$ is a convex continuous function. Now, for condition 1, we can directly evoke Theroem 2 in Stegall (1978); for condition 2, we can directly evoke Corollary 3.11 in Lindenstrauss and Preiss (2003).

step 2 From now, without loss of generality, we may assume that $\mathbb{R}^{J+1} \times \mathcal{X}$ is equipped with the norm $\left(\|r\|^2 + \|x\|_{\mathcal{X}}^2\right)^{\frac{1}{2}}$. Let $\mathcal{X}$ be a Banach space such that every separable subspace $\mathcal{Y}$ has a separable dual space $\mathcal{Y}^*$, $\mathcal{Z} \subset \mathbb{R}^{J+1} \times \mathcal{X}$ be a separable subspace of $\mathbb{R}^{J+1} \times \mathcal{X}$. Then, $\mathcal{X}' = \{x : \exists (r, x) \in \mathcal{Z}\}$ is a subspace of $\mathcal{X}$, $R' = \{r : \exists (r, x) \in \mathcal{Z}\}$ is a subspace of $\mathbb{R}^{J+1}$, and $\mathcal{Z} \subset R' \times \mathcal{X}'$. The separability of $\mathcal{Z}$ implies the separability of $\mathcal{X}'$. Note that the dual space of $R' \times \mathcal{X}'$ isometrically isomorphic to $(R')^* \times (\mathcal{X}')^*$. Obviously, $\mathbb{R}^{J+1}$ is a separable Banach space with separable dual, so by the separability of $(\mathcal{X}')^*$, $(R' \times \mathcal{X}')^*$ is separable. Let $z^* \in \mathcal{Z}^*$, by Hahn-Banach theorem, there exists an element $(r, x)^* \in (R', \mathcal{X}')^*$ such that $(r, x)^* |_{\mathcal{Z}} = z^*$ and $\|z^*\| = \|(r, x)^*\|$. Therefore, with some misuse of language, we can write that $\mathcal{Z}^* \subset (R' \times \mathcal{X}')^*$. Therefore, the separability of $(R' \times \mathcal{X}')^*$ implies the separability of $\mathcal{Z}^*$. In another word, if $\mathcal{Z}$ is a separable subspace of $\mathbb{R}^{J+1} \times \mathcal{X}$, then $\mathcal{Z}^*$ is separable.

step 3 We need to prove that $\gamma(\lambda, g)$ is locally Lipschitz continuous. Note that

$$\gamma(\lambda, g) = \sup_{\phi\in\Phi} \mathbb{E}\left[\sum_j \lambda_j \phi_j g_j\right] \leqslant \sup_{\phi\in\Phi} \mathbb{E}\left[\sum_j |\lambda_j| \, |g_j| \, \phi_j\right] \leqslant \left(\sum_j |\lambda_j|\right) \mathbb{E}\,|g_i| \, .$$

By the continuous embedding assumption $\mathbb{E}\,|g_j| \leqslant C \|g_j\|_{\mathcal{X}}$ for some constant C. Consider an open neighborhood $B_{\lambda,g,r} := B(\lambda, r) \times B(g, r)$ for arbitrarily chosen $(\lambda, g) \in \mathbb{R}^{J+1} \times \mathcal{X}$, where open balls $B(\lambda, r)$ and $B(g, r)$ are taken in $\mathbb{R}^{J+1}$ and $\mathcal{X}$ respectively. Obviously, there exists a constant M such that $\forall \lambda' \in B(\lambda, r)$, $\|\lambda'\| \leqslant M$. Therefore, for all $(\lambda', g') \in B_{\lambda,g,r}$, $v(\lambda', g')$ is upper bounded. Now, we can show that $v(\lambda', g')$ is also locally lower bounded near λ, g. Let $z = g + \rho(g_0 - g)$ such that $g_0 \in \mathcal{X}$, $z \in B(g, r)$ and $\rho > 1$. It is not hard to see such a point z exists. For arbitrarily chosen $\lambda' \in B(\lambda, r)$, consider

$$\mathcal{V} = \left\{ v : v = \left(1 - \frac{1}{\rho}\right) g + \frac{1}{\rho} z, g \in B(g, r) \right\}.$$

$\mathcal{V}$ is equal to the ball in $\mathcal{X}$ with center $g_0 = \left(1 - \frac{1}{\rho}\right) g + \frac{1}{\rho} z$ and radius $\left(1 - \frac{1}{\rho}\right) r$.For all $v \in \mathcal{V}$, $(2g_1 - v) \in \mathcal{V}$, so there is

$$\gamma\left(\lambda', g_0\right) \leqslant \frac{1}{2}\gamma\left(\lambda', v\right) + \frac{1}{2}\gamma\left(\lambda', 2g_1 - v\right)$$

which implies that

$$\gamma\left(\lambda', v\right) \geqslant 2\gamma\left(\lambda', g_1\right) - \gamma\left(\lambda', 2g_1 - v\right) \geqslant 2\gamma\left(\lambda', g_1\right) - M.$$

Since $\gamma(\lambda', g_0)$ is convex with respect to λ' on $\mathbb{R}^{J+1}$, it is locally Lipschitz with respect to λ', see for example Lemma 14.26 in Villani et al. (2009). So we get that $\gamma(\lambda', g)$ is locally bounded on a neighborhood of (λ, g). We may still denoted the bound as M.

step 4 Now, we can prove that $\gamma(\cdot, \cdot)$ is jointly locally Lipschitz. Let $g_1, g_2 \in B(g, r')$, $g_1 \neq g_2$ where r' is taken small enough such that for all $\lambda' \in B(\lambda, r)$, $|\gamma(\lambda', g')| \leqslant M$ for all $g \in B(g, 2r')$.Next, let $w = g_2 + \frac{r'}{d}(g_2 - g_1)$ where $d = \|g_1 - g_2\|_{\mathcal{X}}$. Obviously, $w \in B(g, 2r')$ and $g_2 = \frac{r'}{r'+d} g_1 + \frac{d}{r'+d} w$. By the partial convexity of $\gamma(\cdot, \cdot)$ and let $\lambda' \in B(\lambda, r)$,

$$\gamma\left(\lambda', g_2\right) \leqslant \frac{r'}{r' + d}\gamma\left(\lambda', g_1\right) + \frac{d}{r' + d}\gamma\left(\lambda', w\right).$$

Then, we have

$$\begin{aligned}\gamma\left(\lambda', g_2\right) - \gamma\left(\lambda', g_1\right) &\leqslant \frac{d}{r' + d}\left(\gamma\left(\lambda', w\right) - \gamma\left(\lambda', g_1\right)\right) \leqslant \frac{d}{r'}\left(\gamma\left(\lambda', w\right) - \gamma\left(\lambda', g_1\right)\right) \\ &\leqslant \frac{2d}{r'} M = \frac{2M}{r'} \|g_2 - g_3\|_{\mathcal{X}}.\end{aligned}$$

Therefore, there exists a constant M_1, a neighborhood $\mathcal{N}_g$ of g and a neighborhood $\mathcal{N}_\lambda$ of λ such that for all $\lambda' \in \mathcal{N}_\lambda$, $\gamma(\lambda', \cdot)$ is Lipschitz with the constant M_1 on $\mathcal{N}_g$. Similar, $\gamma(\cdot, g)$ is also locally Lipschitz in the above uniform sense with a constant M_2. Now, in a neighborhood of (λ, g), we have

$$\begin{aligned}\left|\gamma(\lambda', g') - \gamma(\lambda'', g'')\right| &\leqslant \left|\gamma(\lambda', g') - \gamma(\lambda', g'')\right| + \left|\gamma(\lambda', g'') - \gamma(\lambda'', g'')\right| \\ &\leqslant (M_1 + M_2)\left\|(\lambda', g') - (\lambda'', g'')\right\|_{\mathbb{R}^{J+1}\times\mathcal{X}}.\end{aligned}$$

Now, we can directly evoke the Theorem 2.5 of Preiss (1990) to get that $\gamma(\cdot, \cdot)$ is Fréchet differentiable at least on a dense subset of $\mathbb{R}^{J+1} \times \mathcal{X}$.

step 5 In order to calculate the Fréchet derivative of $\gamma(\lambda, g)$ at those Fréchet differentiable points, we use the auxiliary function defined in (18). $\mathbb{E}\left[\sum_j \lambda'_j \phi^*_j g'_j\right]$ is continuous linear by Proposition 3.1.11 in Niculescu and Persson (2018) and thus Fréchet differentiable with respect to (λ, g) with the derivate in the right hand side of (19). Obviously, $e(\lambda', g'; \phi^*)$ is concave and achieves its maximum at (λ, g). Therefore, by Theorem 3.6.11 of Niculescu and Persson (2018), we have $0 \in \nabla^+ e(\lambda', g'; \phi^*)\,|_{(\lambda', g')=(\lambda, g)}$. So, if $\gamma(\lambda', g')$ is Fréchet differentiable at (λ, g), by Propsoition 3.6.9 in Niculescu and Persson (2018),

$$\nabla^+\left(\lambda', g'; \phi^*\right)|_{(\lambda', g')=(\lambda, g)} = \left\{e'\left(\lambda', g'; \phi^*|_{(\lambda', g')=(\lambda, g)}\right)\right\} = \{0\}.$$

Now, the proof ends with direct calculation. ■

The conditions for Fréchet differentiability stated in Theorem 4.2.1 originate from the partial convexity of $\gamma(\lambda, g)$ in either λ or g, and the joint local Lipschitz property of $\gamma(\lambda, g)$ in (λ, g) and make use of the key results Stegall (1978) (Theorem 2), Lindenstrauss and Preiss (2003) (Corollary 3.11) and Preiss (1990) (Theorem 2.5). Namioka and Phelps (1975) attributed a space $\mathcal{X}$ such that every separable subspace $\mathcal{Y}$ of it has a separable dual space $\mathcal{Y}^*$ to Asplund (1968) as an Asplund space.

Theorem 4.2.1 essentially shows that, under mild conditions, a generic point[3] can support taking the Fréchet derivative of $\gamma(\lambda, g)$. In Chernozhukov et al. (2018b), Hadamard differentiability is verified by direct calculations with great details of the partial effect quantile function (called sorted effect therein). We have demonstrated how to reformate their approach in section 3. Compared with Chernozhukov et al. (2018b), the assumptions in Theorem 4.2.1 is not more restrictive in several senses. First, Hadamard differentiability implies continuity (See for example Averbuh and Smoljanov (1968)). Second, it is well known that for a measure space $(\Omega, \mathcal{F}, \mu)$ such that $\mathcal{F}$ is separable or countably generated, and μ is σ-finite, $L_p(\Omega)$ is separable for $1 < p < \infty$, which implies that $L_p(\Omega)$ is separable and reflexive when Ω is a Polish space, $\mathcal{F}$ is the Borel σ-algebra, and μ is a probability measure. Furthermore, when Ω is an open subset of $\mathbb{R}^n$ and $\mathcal{F}$ is the

[3] A generic property is one that holds on a residual set.

σ-algebra of Lebesgue measure, and $\mu = \mathcal{L}_n$ the n-dimensional Lebesgue measure, both $L_p(\Omega)$ and the Sobolev space $W^{k,p}(\Omega)$ is also separable and reflexive. Third, we do not directly make assumptions about the underlying distribution μ, such as requiring compact support or having a continuous density function.

Given convexity and local Lipschitzness, Under the conditions of Theorem 4.2.1, generic Fréchet differentiability follows from convexity and local Lipschitzness. Consequently, Fréchet differentiability (which is stronger than Hadamard differentiability) is a reasonable assumption. The generic Fréchet differentiability property here provides a conceptual justification of the functional differential methodology in Chen et al. (2003) and Chernozhukov et al. (2015). Theorem 4.2.1 also differs from the envelope theorems in Milgrom and Segal (2002). Milgrom and Segal (2002) characterized possibly non-convex problems for finite dimensional relevant parameters, while in this work, Theorem 4.2.1 focuses on a generic property for larger relevant parameter spaces.

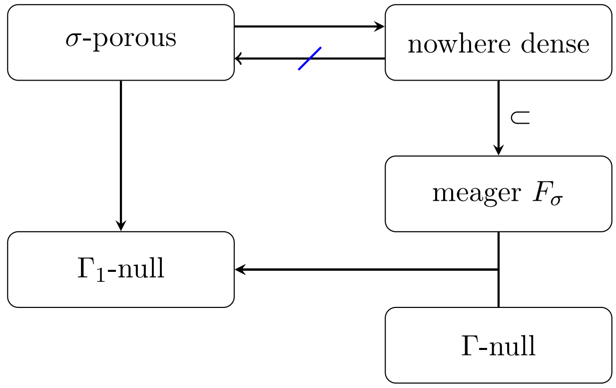
Condition 2 in Theorem 4.2.1 requires that $\mathcal{X}^*$ is separable, which is more restrictive than the Asplund space in condition 1. Theorem 2.4 in Lindenstrauss and Preiss (2003) shows that on $\mathbb{R}^n$, a Γ-null set is equivalent to a set with Lebesgue zero. By the Rademacher Theorem again, given g, $\gamma(\lambda, g)$ is Γ-almost everywhere differentiable with respect to λ.

For a Banach space with a separable dual space, part 1 of Theorem 4.2.1 shows that the set of non Fréchet-differentiable points is contained in a meager F_σ set, while part 2 of Theorem 4.2.1 shows that it is also a Γ-null set. In other words, the set of non Fréchet differentiable points is a Γ-null set, and is contained in a meager F_σ set. However, to the best of our knowledge, there is no general ordering between Γ-null sets and meager F_σ sets. Even in the simplest case $\mathbb{R}$, there are interesting counterexamples. The Smith-Volterra-Cantor set (or called fat Cantor set) is closed and nowhere dense and thus meager F_σ but of positive Lebesgue measure. It is also possible to construct a dense G_δ subset B of $\mathbb{R}$ with zero Lebesgue measure that is not a meager F_σ set (by the Baire theorem). To construct B, enumerate the rational numbers as $\mathbb{Q} = \{q_1, q_2, \ldots\}$, and put $P_n = \bigcup_{j=1}^{\infty} B\left(q_j, 2^{-j}/n\right)$ for positive integers n. Define $B = \bigcap_{n=1}^{\infty} P_n$. Since P_n is open for every positive integer n, we have B is a G_δ set. The set B has zero Lebesgue measure because

$$\mu(B) \leqslant \mu(P_n) \leqslant \sum_{j=1}^{\infty} \mu\left(B\left(q_j, 2^{-j}/n\right)\right) = \sum_{j=1}^{\infty} \frac{2^{-j+1}}{n} = \frac{2}{n} \to 0 \text{ as } n \to \infty.$$

It is worth noting that the concepts of meager F_σ sets (the complement of dense G_δ sets) and Γ-null sets are among a collection of related definitions used to describe the smallness of sets. Part 1 of Theorem 4.2.1 describes a pure topological property of the set of $g \in \mathcal{X}$ at which $\gamma(\lambda, g)$ is Fréchet differentiable, while part 2 of Theorem 4.2.1 is both topological and measure theoretic in nature. Under the conditions of part 2 of Theorem 4.2.1, the set of g at which $\gamma(\lambda, g)$ is not Fréchet differentiable satisfies other smallness properties. For example, on the one hand, by Theorem 1 in Preiss and Zajíček (1984), if $\mathcal{X}$ is a Banach space with separable dual space, then a continuous

Figure 1: Relation between different concepts of small sets



convex function f on $\mathcal{X}$ is Fréchet differentiable outside a σ-porous set. By Theorem 10.4.1 in Lindenstrauss et al. (2012), a σ-porous set is also a Γ_1-null set. On the other hand, by Proposition 5.4.3 in Lindenstrauss et al. (2012), if $A \subset \mathcal{X}$ is a Γ-null and F_σ set, then it is Γ_n null for every n. To recapitulate, σ-porous is a metric space concept, while Γ_n-null is a topological and measure theoretic concept. These relations are illustrated in Figure 1.

Definition 4.2.3. (Porous and σ-porous) A set A in a Banach space $\mathcal{X}$ is called σ-porous if it is a countable union of porous sets. A porous set $E \subset \mathcal{X}$ is such that there exists $0 < c < 1$, and for all $x \in \mathcal{X}$, and for all $\epsilon > 0$, there is a $y \in \mathcal{X}$ satisfying $0 < d(x, y) < \epsilon$ and $B(y, c \cdot d(x, y)) \cap E = \varnothing$.

Definition 4.2.4. (Γ_n-null) Consider $\Gamma_n(\mathcal{X}) = C^1([0,1]^n, \mathcal{X})$ equipped with the $\|\cdot\|_{\leqslant n}$ norm. Then a Borel set $A \subset \mathcal{X}$ is called Γ_n-null if

$$L^n\{t \in [0,1]^n : \gamma(t) \in A\} = 0.$$

for residually many $\gamma \in \Gamma_n(\mathcal{X})$. If A is not Borel, it is also called Γ_n-null if it is contained in a Γ_n-null Borel set.

When λ is given and fixed, a stronger global Lipschitz continuity property can be seen to hold based on $|\max_j\{a_j\} - \max_j\{b_j\}| \leqslant \max_j |a_j - b_j|$:

$$\left|\max_j \lambda_j g_j'(x) - \max_j \lambda_j g_j(x)\right| \leqslant \max_j |\lambda_j| \left|g_j'(x) - g_j(x)\right|.$$

It then follows that

$$|\gamma(\lambda, \hat{g}) - \gamma(\lambda, g)| \leqslant \mathbb{E} \max_j |\lambda_j||\hat{g}_j(X) - g_j(X)| \leqslant \max_j |\lambda_j| \, \|\hat{g} - g\|_{L^1} \leqslant \max_j |\lambda_j| \, \|\hat{g} - g\|_{L^p}, \ p \geqslant 1,$$

where $\hat{g}$ is an estimator of g. Assuming measurability, we have

$$\mathbb{E}_n \left|\gamma\left(\lambda, \hat{g}\right) - \gamma\left(\lambda, g\right)\right| \leqslant \max_j |\lambda_j| \mathbb{E}_n \left\|\hat{g} - g\right\|_{L^p}, \ p \geqslant 1$$

where $\mathbb{E}_n$ applies to the estimation uncertainty in obtaining $\hat{g}(\cdot)$.

4.3 Fast convergence rates

It is also worth noting that our analysis is closely related to the fast learning rates for plug-in classifiers in Devroye et al. (1996) and Audibert and Tsybakov (2007). To illustrate these results consider a weighted version of the population accuracy measure defined as

$$\begin{aligned} \gamma\left(\tau, p\left(\cdot\right), R\right) &= \mathbb{E}\left[\tau Y 1\left(X \in R\right) + \left(1-\tau\right)\left(1-Y\right) 1\left(X \in R^c\right)\right] \\ &= \mathbb{E}\left[\left(1-\tau\right)\left(1-p\left(X\right)\right)\right] + \mathbb{E}\left[\left(p\left(X\right) - \left(1-\tau\right)\right) 1\left(X \in R\right)\right] \end{aligned} \tag{20}$$

Then an optimal choice of R, which is $1\left(p\left(x\right) > \left(1-\tau\right)\right)$, for a given τ leads to

$$\begin{aligned} \gamma\left(\tau, p\left(\cdot\right)\right) &= \max_R \gamma\left(\tau, p\left(\cdot\right), R\right) \\ &= \mathbb{E}\left[\left(1-\tau\right)\left(1-p\left(X\right)\right)\right] + \mathbb{E}\left[\left(p\left(X\right) - \left(1-\tau\right)\right) 1\left(p\left(X\right) > \left(1-\tau\right)\right)\right]. \end{aligned}$$

The welfare regret is defined as the difference between the optimized population welfare and the feasible welfare based on an estimate of the optimal policy $\hat{R} = 1\left(\hat{p}\left(x\right) > 1-\tau\right)$:

$$\begin{aligned} \gamma\left(\tau, p\left(\cdot\right)\right) - \gamma\left(\tau, p\left(\cdot\right), \hat{R}\right) &= \mathbb{E}\left[\left(p\left(X\right) - \left(1-\tau\right)\right)\left(1\left(p\left(X\right) > 1-\tau\right) - 1\left(\hat{p}\left(X\right) > 1-\tau\right)\right)\right] \\ &= \mathbb{E}\left[\left|\left(p\left(X\right) - \left(1-\tau\right)\right)\right| \left|\left(1\left(p\left(X\right) > 1-\tau\right) - 1\left(\hat{p}\left(X\right) > 1-\tau\right)\right)\right|\right]. \end{aligned}$$

The key insight for analyzing convergence speed of above regret is that on the set where $1\left(p\left(x\right) > \left(1-\tau\right)\right) - 1\left(\hat{p}\left(x\right) > \left(1-\tau\right)\right) \neq 0$, or where either $\{p\left(x\right) > 1-\tau, \hat{p}\left(x\right) \leqslant 1-\tau\}$ or $\{p\left(x\right) \leqslant 1-\tau, \hat{p}\left(x\right) > 1-\tau\}$:

$$\left|p\left(x\right) - \left(1-\tau\right)\right| \leqslant \left|p\left(x\right) - \hat{p}\left(x\right)\right|.$$

Therefore the regret can be bounded by,

$$\begin{aligned} \left|\gamma\left(\tau, p\left(\cdot\right)\right) - \gamma\left(\tau, p\left(\cdot\right), \hat{R}\right)\right| &\leqslant \mathbb{E}\left[\left|p\left(X\right) - \hat{p}\left(X\right)\right| \left|\left(1\left(p\left(X\right) > 1-\tau\right) - 1\left(\hat{p}\left(X\right) > 1-\tau\right)\right)\right|\right] \\ &\leqslant \sup_{x \in \mathcal{X}} \left|p\left(x\right) - \hat{p}\left(x\right)\right| \mathbb{E}\left|\left(1\left(p\left(X\right) > 1-\tau\right) - 1\left(\hat{p}\left(X\right) > 1-\tau\right)\right)\right| \end{aligned}$$

As in Chen et al. (2003), for $\delta_n = \sup_{x \in \mathcal{X}} \left|p\left(x\right) - \hat{p}\left(x\right)\right|$,

$$\left|\left(1\left(p\left(x\right) > 1-\tau\right) - 1\left(\hat{p}\left(x\right) > 1-\tau\right)\right)\right| \leqslant 1\left(1-\tau-\delta_n \leqslant p\left(x\right) \leqslant 1-\tau+\delta_n\right).$$

As such the regret can be further bounded by

$$\left|\gamma\left(\tau,p\left(\cdot\right)\right)-\gamma\left(\tau,p\left(\cdot\right),\hat{R}\right)\right|\leqslant\delta_n\mathbb{P}\left(1-\tau-\delta_n\leqslant p\left(X\right)\leqslant 1-\tau+\delta_n\right)$$

Under the conditions in section 3 (see for example Proposition 3.2.3) where $p\left(x\right)$ has a bounded density at $1-\tau$, for a constant C,

$$\mathbb{P}\left(1-\tau-\delta_n\leqslant p\left(X\right)\leqslant 1-\tau+\delta_n\right)\leqslant C\delta_n.$$

In conclusion, $\left|\gamma\left(\tau,p\left(\cdot\right)\right)-\gamma\left(\tau,p\left(\cdot\right),\hat{R}\right)\right|\leqslant C\delta_n^2$. Many nonparametric estimators $\hat{p}\left(x\right)$ achieves $\delta_n=O_{\mathbb{P}}\left(n^{-\frac{\beta}{2\beta+d}}\log n\right)$ where d is the dimension of x, and β a smoothness parameter, implying that

$$\left|\gamma\left(\tau,p\left(\cdot\right)\right)-\gamma\left(\tau,p\left(\cdot\right),\hat{R}\right)\right|=O_{\mathbb{P}}\left(n^{-\frac{2\beta}{2\beta+d}}\log^2 n\right)$$

and whenever $\beta>d/2$,

$$\left|\gamma\left(\tau,p\left(\cdot\right)\right)-\gamma\left(\tau,p\left(\cdot\right),\hat{R}\right)\right|=O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).$$

Note that in the above, we provide primitive conditions that offer the case of $\alpha\geqslant 1$ for the *margin* assumption (MA) in Mammen and Tsybakov (1999), Tsybakov (2004), Boucheron et al. (2005), Audibert and Tsybakov (2007), Kitagawa and Tetenov (2018) and Luedtke and Chambaz (2020).

The misclassification error rate in Audibert and Tsybakov (2007) is the most important case of the weighted population accuracy (20) where $\tau=\frac{1}{2}$, since

$$\begin{aligned}\mathbb{P}\left(Y\neq 1\left(X\in R\right)\right)&=1-\mathbb{P}\left(Y=1\left(X\in R\right)\right)\\&=1-2\left[\frac{1}{2}\mathbb{E}Y1\left(X\in R\right)+\frac{1}{2}\mathbb{E}\left(1-Y\right)1\left(X\in R^c\right)\right]\\&=1-2\gamma\left(\frac{1}{2},p\left(\cdot\right),R\right),\end{aligned}$$

where $p\left(x\right)=\mathbb{E}\left(Y|X=x\right)$.

Recall that in a general discrete allocation problem, we define welfare under policy as

$$\gamma\left(\lambda,g,\phi\right)=\mathbb{E}\left[\sum_{j=0}^{J}\lambda_j\phi_j\left(X\right)Y_j\right]=\mathbb{E}\left[\sum_{j=0}^{J}\lambda_j\phi_j\left(X\right)g_j\left(X\right)\right],$$

where $g_j\left(x\right)=\mathbb{E}\left(Y_j|X=x\right)$ and $\phi_j\left(x\right)\geqslant 0,\forall j$, $\sum_{j=0}^{J}\phi_j\left(x\right)=1$. Then an optimal choice of $\phi\left(\cdot\right)$, which be given by (2), for a given λ leads to the social welfare potention function (1). The feasible analog based on an estimate $\hat{g}(x)$ should then satisfy

$$\hat{\phi}_j(X) = \begin{cases} 1, & \text{if } \lambda_j \hat{g}_j(X) > \lambda_l \hat{g}_j(X), \ \forall l \neq j, \\ \\ 0, & \text{if } \lambda_j \hat{g}_j(X) < \lambda_l \hat{g}_j(X), \ \exists l \neq j. \end{cases}$$

If there exists some l such that $\lambda_j \hat{g}_j(X) = \lambda_l \hat{g}_l(X)$, then $\hat{\phi}(\cdot)$ can be divided among all the maximal indexes. In the following we adopt a convention of allocating the optimal $\phi^*(x)$ and $\hat{\phi}(x)$ such that all the weight among the maximal indexes is allocated to the smallest index member:

$$\phi_j^*(x) = \prod_{l=j+1}^{J} 1\left(\lambda_j g_j(x) \geqslant \lambda_l g_l(x)\right) \prod_{l=0}^{j-1} 1\left(\lambda_j g_j(x) > \lambda_l g_l(x)\right),$$

$$\hat{\phi}_j(x) = \prod_{l=j+1}^{J} 1\left(\lambda_j \hat{g}_j(x) \geqslant \lambda_l \hat{g}_l(x)\right) \prod_{l=0}^{j-1} 1\left(\lambda_j \hat{g}_j(x) > \lambda_l \hat{g}_l(x)\right).$$

The welfare regret is defined as the difference between the social welfare function and its feasible version based on the plug-in estimate of the optimal policy:

$$\gamma(\lambda, g, \phi^*) - \gamma\left(\lambda, g, \hat{\phi}\right) = \sum_{j=0}^{J} \mathbb{E}\left[\lambda_j g_j(X)\left(\phi_j^*(X) - \hat{\phi}_j(X)\right)\right]$$

Note that by definition, for each x, $\phi_j^*(x)$ is either 0 or 1 for each $j = 0, \ldots, J$. Furthermore, one and only one of $\phi_j^*(x)$ out of $j = 0, \ldots, J$ takes the value 1. The same is also true for $\hat{\phi}_j(x)$ for each x. Therefore we can write

$$\begin{aligned}
&\gamma(\lambda, g, \phi^*) - \gamma\left(\lambda, g, \hat{\phi}\right) \\
&= \mathbb{E}\left[\sum_{i \neq j} \left(\lambda_i g_i(X) - \lambda_j g_j(X)\right) \phi_i^*(X) \hat{\phi}_j(X)\right] \\
&\leqslant \mathbb{E}\left[\sum_{i \neq j} \left|\lambda_i g_i(X) - \lambda_j g_j(X) - \left(\lambda_i \hat{g}_i(X) - \lambda_j \hat{g}_j(X)\right)\right| \phi_i^*(X) \hat{\phi}_j(X)\right] \\
&= \mathbb{E}\left[\sum_{i \neq j} \left|\lambda_i g_i(X) - \lambda_j g_j(X) - \left(\lambda_i \hat{g}_i(X) - \lambda_j \hat{g}_j(X)\right)\right| 1\left(\phi_i^*(X) = \hat{\phi}_j(X) = 1\right)\right] \\
&\leqslant \mathbb{E}\left[\sum_{i \neq j} \left|\lambda_i g_i(X) - \lambda_j g_j(X) - \left(\lambda_i \hat{g}_i(X) - \lambda_j \hat{g}_j(X)\right)\right| 1\left(\phi^*(X) \neq \hat{\phi}(X)\right)\right] \\
&\leqslant \mathbb{E}\left[\sum_{i \neq j} \left|\lambda_i g_i(X) - \lambda_j g_j(X) - \left(\lambda_i \hat{g}_i(X) - \lambda_j \hat{g}_j(X)\right)\right| \sum_{k \neq l} \left|1_{kl}^* - \hat{1}_{kl}\right|\right]
\end{aligned} \tag{21}$$

where we define, for $k \neq l, k, l = 0, \ldots, J$,

$$1^*_{kl}(x) = 1(k < l)\, 1(\lambda_k g_k(x) \geqslant \lambda_l g_l(x)) + 1(k > l)\, 1(\lambda_k g_k(x) > \lambda_l g_l(x)).$$

and

$$\hat{1}_{kl}(x) = 1(k < l)\, 1(\lambda_k \hat{g}_k(x) \geqslant \lambda_l \hat{g}_l(x)) + 1(k > l)\, 1(\lambda_k \hat{g}_k(x) > \lambda_l \hat{g}_l(x)).$$

As previously, following Chen et al. (2003), for $\delta_n = \sup_{x \in \mathcal{X}} \max_{j=0,\ldots,J} |\lambda_j g_j(x) - \lambda_j \hat{g}_j(x)|$,

$$|1^*_{kl} - \hat{1}_{kl}| \leqslant 1(-2\delta_n \leqslant \lambda_k g_k(x) - \lambda_l g_l(x) \leqslant 2\delta_n). \tag{22}$$

Then we can further bound the regret as

$$\gamma(\lambda, g, \phi^*) - \gamma\left(\lambda, g, \hat{\phi}\right) \leqslant 2(J+1)^2 \delta_n \sum_{k \neq l} \mathbb{P}(-2\delta_n \leqslant \lambda_k g_k(X) - \lambda_l g_l(X) \leqslant 2\delta_n).$$

Under the conditions in section 3 where for each pair $k \neq l$, $\lambda_k g_k(x) - \lambda_l g_l(x)$ has a bounded density at zero, $\mathbb{P}(-2\delta_n \leqslant \lambda_k g_k(X) - \lambda_l g_l(X) \leqslant 2\delta_n) \leqslant C\delta_n$, leading to

$$\gamma(\lambda, g, \phi^*) - \gamma\left(\lambda, g, \hat{\phi}\right) \leqslant C\delta_n^2.$$

Given nonparametric estimates of $\hat{g}_j(x), j = 0, \ldots, J$ achieving $\delta_n = O_{\mathbb{P}}\left(n^{-\frac{\beta}{2\beta+d}} \log n\right)$, there is

$$\left|\gamma(\lambda, g) - \gamma\left(\lambda, g, \hat{\phi}\right)\right| = O_{\mathbb{P}}\left(n^{-\frac{2\beta}{2\beta+d}} \log^2 n\right)$$

and whenever $\beta > d/2$, $|\gamma(\lambda, g) - \gamma\left(\lambda, g, \hat{\phi}\right)| = o_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$.

The unknown social welfare potential function $\beta_0 = \gamma(\lambda, g) = \mathbb{E} \max_{j=0,\ldots,J} \lambda_j g_j(x)$ can be estimated by its sample analog $\hat{\gamma}(\lambda, \hat{g}(\cdot)) = \frac{1}{n'} \sum_{i=1}^{n'} \max_{j=0,\ldots,J} \lambda_j \hat{g}_j(X_i)$. Following Chernozhukov et al. (2018a) and Chernozhukov et al. (2022), we consider a Neyman-orthogonal debiased version as, where with $Y_i = \sum_{j=0}^{J} D_{ij} Y_{ij}$,

$$\hat{\beta} = \frac{1}{n'} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \hat{\phi}_j(X_i) \left[\hat{g}_j(X_i) + \frac{D_{ij}}{\hat{p}_j(X_i)} (Y_i - \hat{g}_j(X_i))\right]$$

where $D_{ij} = 1$ if $D_i = j$, $p_j(x_i) = \mathbb{P}(D_{ij} = 1 | X_i = x_i)$, $g_j(x_i) = \mathbb{E}(Y_j | D_{ij} = 1, X_i = x_i)$. We should note that Luedtke and Van Der Laan (2016b) and Luedtke and Chambaz (2020) utilize a formula similar to the one above for the binary case under empirical process settings. In a split sample scheme, $\hat{p}(\cdot)$ and $\hat{g}(\cdot)$ that are applied to each X_i are obtained in a different subsample that excludes $\{X_i\}_{i=1}^{n'}$. The goal is to show that

$$\sqrt{n'}\left(\hat{\beta}-\beta_0\right)=\frac{1}{\sqrt{n'}}\sum_{i=1}^{n'}\left(\sum_{j=0}^{J}\lambda_j\phi_j^*\left(X_i\right)\left[g_j\left(X_i\right)+\frac{D_{ij}}{p_j\left(X_i\right)}\left(Y_i-g_j\left(X_i\right)\right)\right]-\beta_0\right)+o_{\mathbb{P}}\left(1\right).$$

This will follow if we can show that $\Delta=o_{\mathbb{P}}\left(1\right)$, where

$$\begin{aligned}\Delta=\frac{1}{\sqrt{n'}}\sum_{i=1}^{n'}\sum_{j=0}^{J}\Bigg(&\lambda_j\hat{\phi}_j\left(X_i\right)\left[\hat{g}_j\left(X_i\right)+\frac{D_{ij}}{\hat{p}_j\left(X_i\right)}\left(Y_i-\hat{g}_j\left(X_i\right)\right)\right]\\&-\lambda_j\phi_j^*\left(X_i\right)\left[g_j\left(X_i\right)+\frac{D_{ij}}{p_j\left(X_i\right)}\left(Y_i-g_j\left(X_i\right)\right)\right]\Bigg).\end{aligned}$$

For this purpose we decompose $\Delta=\Delta_1+\Delta_2$, where

$$\Delta_1=\frac{1}{\sqrt{n'}}\sum_{i=1}^{n'}\sum_{j=0}^{J}\lambda_j\hat{\phi}_j\left(X_i\right)\left[\hat{g}_j\left(X_i\right)+\frac{D_{ij}}{\hat{p}_j\left(X_i\right)}\left(Y_i-\hat{g}_j\left(X_i\right)\right)-g_j\left(X_i\right)-\frac{D_{ij}}{p_j\left(X_i\right)}\left(Y_i-g_j\left(X_i\right)\right)\right]$$

and

$$\Delta_2=\frac{1}{\sqrt{n'}}\sum_{i=1}^{n'}\sum_{j=0}^{J}\lambda_j\left(\hat{\phi}_j\left(X_i\right)-\phi_j^*\left(X_i\right)\right)\left[g_j\left(X_i\right)+\frac{D_{ij}}{p_j\left(X_i\right)}\left(Y_i-g_j\left(X_i\right)\right)\right].$$

Also define a linearized approximation of Δ_1 as

$$\Delta_3=\frac{1}{\sqrt{n'}}\sum_{i=1}^{n'}\sum_{j=0}^{J}\lambda_j\hat{\phi}_j\left(X_i\right)\left[\left(1-\frac{D_{ij}}{p_j\left(X_i\right)}\right)\left(\hat{g}_j\left(X_i\right)-g_j\left(X_i\right)\right)-\frac{D_{ij}}{p_j^2\left(X_i\right)}\left(Y_i-g_j\left(X_i\right)\right)\left(\hat{p}_j\left(X_i\right)-p_j\left(X_i\right)\right)\right].$$

Then we can write, when $p\left(x\right)$ and $\hat{p}\left(x\right)$ is bounded away from zero,

$$\left|\Delta_1-\Delta_3\right|\leqslant C\frac{1}{\sqrt{n'}}\sum_{i=1}^{n'}\sum_{j=0}^{J}\left(\left|\hat{g}_j\left(X_i\right)-g_j\left(X_i\right)\right|^2+\left|\hat{p}_j\left(X_i\right)-p_j\left(X_i\right)\right|^2\right).$$

Under suitable conditions, $\left|\Delta_1-\Delta_3\right|=o_{\mathbb{P}}\left(1\right)$. To see that $\Delta_3=o_{\mathbb{P}}\left(1\right)$, note that by the split sample scheme, conditional on the estimate $\hat{p}\left(\cdot\right)$ and $\hat{g}\left(\cdot\right)$, $\mathbb{E}\Delta_3=0$,

$$\begin{aligned}Var\left(\Delta_3\right)&\leqslant 2\sum_{j=0}^{J}\lambda_j^2\Bigg(\mathbb{E}\left[\frac{1-p_j\left(X_i\right)}{p_j\left(X_i\right)}\left(\hat{g}\left(X_i\right)-g\left(X_i\right)\right)^2\hat{\phi}_j\left(X_i\right)^2\right]\\&\quad+\mathbb{E}\left[\hat{\phi}_j\left(X_i\right)^2Var\left(Y_{ij}|X_i,D_{ij}=1\right)\frac{1}{p_j^3\left(X_i\right)}\left(\hat{p}\left(X_i\right)-p\left(X_i\right)\right)^2\right]\Bigg)\\&\leqslant C\left(\mathbb{E}\left(\hat{g}\left(X_i\right)-g\left(X_i\right)\right)^2+\mathbb{E}\left(\hat{p}\left(X_i\right)-p\left(X_i\right)\right)^2\right).\end{aligned}$$

Next we decompose $\Delta_2 = \Delta_2^1 + \Delta_2^2$, where

$$\Delta_2^1 = \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \left(\hat{\phi}_j (X_i) - \phi_j^* (X_i) \right) \frac{D_{ij}}{p_j (X_i)} (Y_i - g_j (X_i)) .$$

and

$$\Delta_2^2 = \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \left(\hat{\phi}_j (X_i) - \phi_j^* (X_i) \right) g_j (X_i) .$$

To see that $\Delta_2^1 = o_{\mathbb{P}} (1)$, note that $\mathbb{E}\Delta_2^1 = 0$, and

$$Var\left(\Delta_2^1\right) \leqslant \sum_{j=0}^{J} \lambda_j^2 \mathbb{E}\left[\left(\hat{\phi}_j (X_i) - \phi_j^* (X_i) \right)^2 \frac{1}{p_j (X_i)} Var\left(Y_j | X_i, D_{ij} = 1 \right) \right] \leqslant C \sum_{j=0}^{J} \mathbb{E} \left(\hat{\phi}_j (X_i) - \phi_j^* (X_i) \right)^2,$$

which is bounded using the same method as in the following. Using arguments analogous to (21) and (22), we can further bound $|\Delta_2^2|$

$$\begin{aligned}
-\Delta_2^2 &= \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{j=0}^{J} \lambda_j \left(\phi_j^* (X_i) - \hat{\phi}_j (X_i) \right) g_j (X_i) \\
&\leqslant \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{p \neq q} \left| \lambda_p g_p (X_i) - \lambda_q g_q (X_i) - (\lambda_p \hat{g}_p (X_i) - \lambda_q \hat{g}_q (X_i)) \right| \phi_p^* (X_i) \hat{\phi}_q (X_i) \\
&\leqslant \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{p \neq q} \left[\left| \lambda_p g_p (X_i) - \lambda_q g_q (X_i) - (\lambda_p \hat{g}_p (X_i) - \lambda_q \hat{g}_q (X_i)) \right| 1 \left(\phi^* (X_i) \neq \hat{\phi} (X_i) \right) \right] \\
&\leqslant \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{p \neq q} \left[\left| \lambda_p g_p (X_i) - \lambda_q g_q (X_i) - (\lambda_p \hat{g}_p (X_i) - \lambda_q \hat{g}_q (X_i)) \right| \sum_{k \neq l} \left| 1_{kl}^* (X_i) - \hat{1}_{kl} (X_i) \right| \right] \\
&\leqslant 2 (J+1)^2 \delta_n \frac{1}{\sqrt{n'}} \sum_{i=1}^{n'} \sum_{p \neq q} \left| 1_{kl}^* (X_i) - \hat{1}_{kl} (X_i) \right|
\end{aligned}$$

Given that $-\Delta_2^2 \geqslant 0$, to show that $\Delta_2^2 = o_{\mathbb{P}} (1)$, it suffices to check that

$$\begin{aligned}
-\mathbb{E}\Delta_2^2 &\leqslant \mathbb{E} C \delta_n \sqrt{n'} \sum_{p \neq q} \left| 1_{pq}^* (X_i) - \hat{1}_{pq} (X_i) \right| \\
&\leqslant C \sqrt{n'} \mathbb{E} \delta_n \sum_{p \neq q} \mathbb{P} \left(-2\delta_n \leqslant \lambda_p g_p^* (X) - \lambda_q g_q^* (X) \leqslant 2\delta_n \right) \\
&= C \sqrt{n'} \mathbb{E} \delta_n^2 = o (1),
\end{aligned}$$

where we implicitly require that $\lim_{n \to \infty} \frac{n'}{n} = \rho$, $0 < \rho < \infty$. We should also note that in the above, we somehow misuse the notation C. They should be understood just as *some constant*, and they are not necessarily equal to each other.

5 Conclusion

In this paper, we explore a functional differentiability approach for a class of statistical optimal allocation problem. Inspired by Chernozhukov et al. (2018b), Hadamard differentiability is facilitated by a study of the general properties of the sorting operator. Our derivation depends indispensably on the concept of Hausdorff measure and the area and coarea integration formula from geometric measure theory. Based on our general Hadamard differentiability results, we have showed in subsection 3.3 that both the asymptotic properties of the value function process of the binary constrained optimal allocation problem and the two-step ROC curve estimator can be directly derived from the functional delta method. When the first step propensity score estimator is correctly specified, a computationally feasible bootstrap procedure is also validated for the two-step ROC estimator. More surprisingly, utilizing deep geometric functional analysis results on convex and local Lipschitz functional, we can demonstrate generic Fréchet differentiability for the social welfare potential function. These intriguing results further motivate us to deliberate on estimation methods that take into consideration the first order term of the social welfare potential function. Combining techniques from the literature of nonsmooth method of moment, plug-in classification and the recent development of double / debiased machine learning, we provide a debiased estimator of the social welfare potential function. Here, the conditions required for Hadamard differentiability validate the margin assumption, which leads to a faster convergence rate.

We deliberately avoid aiming for more general but also more complex conditions in the application subsections 3.3 and 4.3, in order to highlight our concise methodology. Besides the details, a key message we would like to convey in this paper is that the value functions of the optimal allocation problems are usually more *regular* than expected. Much can be done by taking advantage of the regularity property. Extensions and generalizations of this paper, including semiparametric efficiency and multiclassification ROC surfaces, are studied in follow-up work by Feng et al. (2024).

6 Acknowledgments

This paper is developed based on the ideas from an earlier work-in-progress project circulated under the title “Decision Making with Machine Learning and ROC Curves”. The previous work-in-progress project latter has been divided into several manuscripts. We acknowledge helpful and encouraging conservations with collaborators and colleagues including Timothy Armstrong, Xiaohong Chen, Hongjun Li, Jessie Li, Thomas MaCurdy. We are most grateful to Yuhao Xue for enlightening us with his knowledge of geometry measure theory. We acknowledge funding support from the National Science Foundation (SES 1658950 to Han Hong).

References

Agrawal, Ajay, Joshua S Gans, and Avi Goldfarb, “Artificial intelligence: the ambiguous labor market impact of automating prediction,” *Journal of Economic Perspectives*, 2019, *33* (2), 31–50.

Aliprantis, Charalambos D and Kim C Border, *Infinite dimensional analysis, A Hitchhiker's Guide*, Springer, 2006.

Armstrong, Timothy B and Shu Shen, "Inference on optimal treatment assignments," *The Japanese Economic Review*, 2023, *74* (4), 471–500.

Asplund, Edgar, "Fréchet differentiability of convex functions," *Acta Mathematica*, 1968, *121*, 31–47.

Athey, Susan and Stefan Wager, "Policy learning with observational data," *Econometrica*, 2021, *89* (1), 133–161.

Audibert, Jean-Yves and Alexandre B Tsybakov, "Fast learning rates for plug-in classifiers," *Annals of statistics*, 2007, *35* (2), 608–633.

Averbuh, VI and OG Smoljanov, "Različnye opredelenija proizvodnoi v lineinhy topologičeskih prostranstvah (Different definitions of derivative in linear topological spaces)," *Uspehi Mat. Nauk*, 1968, *23*, 67.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, "Artificial intelligence, firm growth, and product innovation," *Journal of Financial Economics*, 2024, *151*, 103745.

Ben-Michael, Eli, Kosuke Imai, and Zhichao Jiang, "Policy Learning with Asymmetric Counterfactual Utilities," *Journal of the American Statistical Association*, 2024, pp. 1–14.

Bertail, Patrice, Stéphan Clémençcon, and Nicolas Vayatis, "On bootstrapping the ROC curve," *Advances in Neural Information Processing Systems*, 2008, *21.*

Bhattacharya, Debopam and Pascaline Dupas, "Inferring welfare maximizing treatment assignment under budget constraints," *Journal of Econometrics*, 2012, *167* (1), 168–196.

Bolte, Jérôme and Edouard Pauwels, "Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning," *Mathematical Programming*, 2021, *188*, 19–51.

Boucheron, Stéphane, Olivier Bousquet, and Gábor Lugosi, "Theory of classification: A survey of some recent advances," *ESAIM: probability and statistics*, 2005, *9*, 323–375.

Chen, Xiaohong, Oliver Linton, and Ingrid Van Keilegom, "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 2003, *71* (5), 1591–1608.

Chernozhukov, Victor, Christian Hansen, and Martin Spindler, "Valid post-selection and post-regularization inference: An elementary, general approach," *Annu. Rev. Econ.*, 2015, *7* (1), 649–688.

_ , Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/debiased machine learning for treatment and structural parameters," *Econometrics Journal*, 2018, *21*, C1–C68.

_ , Iván Fernández-Val, and Alfred Galichon, "Quantile and probability curves without crossing," *Econometrica*, 2010, *78* (3), 1093–1125.

_ , _ , and Ye Luo, "The sorted effects method: discovering heterogeneous effects beyond their averages," *Econometrica*, 2018, *86* (6), 1911–1938.

_ , Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins, "Locally robust semiparametric estimation," *Econometrica*, 2022, *90* (4), 1501–1535.

Clarke, Francis, "On the inverse function theorem," *Pacific Journal of Mathematics*, 1976, *64* (1), 97–102.

Clarke, Frank H, "Generalized gradients and applications," *Transactions of the American Mathematical Society*, 1975, *205*, 247–262.

Debreu, Gerard, "Economies with a finite set of equilibria," *Econometrica: Journal of the Econometric Society*, 1970, pp. 387–392.

Dellacherie, Claude and Paul-André Meyer, *Probabilities and potential, A*, Vol. 29 of *North-Holland Mathematics Studies*, North-Holland, 1978.

Devroye, Luc, László Györfi, and Gábor Lugosi, *A Probabilistic Theory of Pattern Recognition*, Vol. 31 of *Stochastic Modelling and Applied Probability*, Springer New York, 1996.

Evans, Lawrence C and Ronald F Garzepy, *Measure theory and fine properties of functions, revised edition*, Chapman and Hall/CRC, 2015.

Falconer, Kenneth J, *The geometry of fractal sets*, Vol. 85 of *Cambridge Tracts in Mathematics*, Cambridge University Press, 1986.

Fang, Zheng and Andres Santos, "Inference on directionally differentiable functions," *The Review of Economic Studies*, 2019, *86* (1), 377–412.

Federer, Herbert, "Surface area. II," *Transactions of the American Mathematical Society*, 1944, *55*, 438–456.

_ , "Curvature measures," *Transactions of the American Mathematical Society*, 1959, *93* (3), 418–491.

_ , "Geometric Measure Theory," *Springer*, 1969.

Feng, Kai, Han Hong, Jessie Li, Ke Tang, and Jingyuan Wang, "Statistical Inference of Optimal Allocations II: ROC Analysis," 2024. Working Paper.

Figalli, Alessio, "A simple proof of the Morse-Sard theorem in Sobolev spaces," *Proceedings of the American Mathematical Society*, 2008, *136* (10), 3675–3681.

Grigoryan, Alexander, *Heat kernel and analysis on manifolds*, Vol. 47 of *AMS/IP Studies in Advanced Mathematics*, American Mathematical Soc., 2009.

Hall, Peter, Rob J Hyndman, and Yanan Fan, "Nonparametric confidence intervals for receiver operating characteristic curves," *Biometrika*, 2004, *91* (3), 743–750.

Hirano, Keisuke and Jack R Porter, "Asymptotics for statistical treatment rules," *Econometrica*, 2009, *77* (5), 1683–1701.

Hiriart-Urruty, Jean-Baptiste, "Tangent cones, generalized gradients and mathematical programming in Banach spaces," *Mathematics of Operations Research*, 1979, *4* (1), 79–97.

Hsieh, Fushing and Bruce W Turnbull, "Nonparametric and semiparametric estimation of the receiver operating characteristic curve," *The Annals of Statistics*, 1996, *24* (1), 25–40.

Jerrard, Robert L., "Lecture note for Geometric Measure Theory," 2013. University of Toronto, http://www.math.toronto.edu/rjerrard/1501/gmt.html.

Kim, Jeankyung and David Pollard, "Cube root asymptotics," *The Annals of Statistics*, 1990, *18* (1), 191–219.

Kitagawa, Toru and Aleksey Tetenov, "Who should be treated? empirical welfare maximization methods for treatment choice," *Econometrica*, 2018, *86* (2), 591–616.

_ **and Guanyi Wang**, "Individualized Treatment Allocation in Sequential Network Games," *arXiv preprint arXiv:2302.05747*, 2023.

— **and** — , "Who should get vaccinated? Individualized allocation of vaccines over SIR network," *Journal of Econometrics*, 2023, *232* (1), 109–131.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 2018, *133* (1), 237–293.

Lehmann, Erich Leo and Joseph P Romano, *Testing statistical hypotheses*, 4 ed., Vol. 3 of *Springer Texts in Statistics*, Springer, 2022.

Li, Gang, Ram C Tiwari, and Martin T Wells, "Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves," *Biometrika*, 1999, *86* (3), 487–502.

Lindenstrauss, Joram and David Preiss, "On Fréchet differentiability of Lipschitz maps between Banach spaces," *Annals of Mathematics*, 2003, pp. 257–288.

— , — , and Jaroslav Tišer, *Fréchet Differentiability of Lipschitz Functions and Porous Sets in Banach Spaces*, Vol. 179 of *Annals of Mathematics Studies*, Princeton University Press, 2012.

Lloyd, Chris J, "Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems," *Journal of the American Statistical Association*, 1998, *93* (444), 1356–1364.

Lu, Xuguang, "Lecture Notes for Mathematical Analysis," September 2019. Tsinghua University.

Luckett, Daniel J, Eric B Laber, Samer S El-Kamary, Cheng Fan, Ravi Jhaveri, Charles M Perou, Fatma M Shebl, and Michael R Kosorok, "Receiver operating characteristic curves and confidence bands for support vector machines," *Biometrics*, 2021, *77* (4), 1422–1430.

Luedtke, Alex and Antoine Chambaz, "Performance guarantees for policy learning," *Annales de l'IHP Probabilites et statistiques*, 2020, *56* (3), 2162.

Luedtke, Alexander R and Mark J van der Laan, "Optimal individualized treatments in resource-limited settings," *The International Journal of Biostatistics*, 2016, *12* (1), 283–303.

— **and Mark J Van Der Laan**, "Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy," *Annals of statistics*, 2016, *44* (2), 713.

Mammen, Enno and Alexandre B Tsybakov, "Smooth discrimination analysis," *The Annals of Statistics*, 1999, *27* (6), 1808–1829.

Manski, Charles F, "Statistical treatment rules for heterogeneous populations," *Econometrica*, 2004, *72* (4), 1221–1246.

Mbakop, Eric and Max Tabord-Meehan, "Model selection for treatment choice: Penalized welfare maximization," *Econometrica*, 2021, *89* (2), 825–848.

Milgrom, Paul and Ilya Segal, "Envelope theorems for arbitrary choice sets," *Econometrica*, 2002, *70* (2), 583–601.

Namioka, Isaac and Robert R Phelps, "Banach spaces which are Asplund spaces," *Duke Mathematical Journal*, 1975, *42* (4), 735–750.

Niculescu, Constantin P and Lars-Erik Persson, *Convex Functions and Their Applications: A Contemporary Approach*, 2 ed., Springer, 2018.

Preiss, D and L Zajíček, "Fréchet differentiation of convex functions in a Banach space with a separable dual," *Proceedings of the American Mathematical Society*, 1984, *91* (2), 202–204.

Preiss, David, "Differentiability of Lipschitz functions on Banach spaces," *Journal of Functional Analysis*, 1990, *91* (2), 312–345.

Qian, Min and Susan A Murphy, "Performance guarantees for individualized treatment rules," *Annals of statistics*, 2011, *39* (2), 1180.

Rai, Yoshiyasu, "Statistical inference for treatment assignment policies," 2018. Unpublished Manuscript.

Resnick, Sidney I, *Extreme values, regular variation, and point processes*, Vol. 4 of *Springer Series in Operations Research and Financial Engineering*, Springer Science & Business Media, 1987.

Rockafellar, R Tyrrell and Roger J-B Wets, *Variational analysis*, Vol. 317 of *Grundlehren der mathematischen Wissenschaften*, Springer Science & Business Media, 2009.

Sasaki, Yuya, "What do quantile regressions identify for general structural functions?," *Econometric Theory*, 2015, *31* (5), 1102–1116.

Shapiro, Alexander, "On concepts of directional differentiability," *Journal of Optimization Theory and Applications*, 1990, *66* (3), 477–487.

Sherman, Robert P, "The limiting distribution of the maximum rank correlation estimator," *Econometrica: Journal of the Econometric Society*, 1993, pp. 123–137.

Smale, S, "An Infinite Dimensional Version of Sard's Theorem," *American Journal of Mathematics*, 1965, *87* (4), 861–866.

Stegall, Charles, "The duality between Asplund spaces and spaces with the Radon-Nikodym property," *Israel Journal of Mathematics*, 1978, *29*, 408–412.

Stoye, Jörg, "Minimax regret treatment choice with finite samples," *Journal of Econometrics*, 2009, *151* (1), 70–81.

Swaminathan, Adith and Thorsten Joachims, "Batch learning from logged bandit feedback through counterfactual risk minimization," *The Journal of Machine Learning Research*, 2015, *16* (1), 1731–1755.

Tetenov, Aleksey, "Statistical treatment choice based on asymmetric minimax regret criteria," *Journal of Econometrics*, 2012, *166* (1), 157–165.

Tsybakov, Alexander B, "Optimal aggregation of classifiers in statistical learning," *The Annals of Statistics*, 2004, *32* (1), 135–166.

van der Vaart, AW and Jon A Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*, 2 ed., Springer Nature, 2023.

Villani, Cédric et al., *Optimal transport: old and new*, Vol. 338 of *Grundlehren der mathematischen Wissenschaften*, Springer, 2009.

Zhao, Yingqi, Donglin Zeng, A John Rush, and Michael R Kosorok, "Estimating individualized treatment rules using outcome weighted learning," *Journal of the American Statistical Association*, 2012, *107* (499), 1106–1118.

Zhou, Xin, Nicole Mayer-Hamblett, Umer Khan, and Michael R Kosorok, "Residual weighted learning for estimating individualized treatment rules," *Journal of the American Statistical Association*, 2017, *112* (517), 169–187.

Zhou, Zhengyuan, Susan Athey, and Stefan Wager, "Offline multi-action policy learning: Generalization and optimization," *Operations Research*, 2023, *71* (1), 148–183.

T Technical addendum

T.1 Supplementary definitions and results

Theorem T.1.1. *Carathéodory criterion*
Let μ^* be an outer measure on a set O. A subset $E \subset O$ is said to be μ^*-measurable if

$$\mu^*(T) = \mu^*(T \cap E) + \mu^*(T \cap E^c) \quad \forall T \subset O.$$

Let $\mathcal{M}$ be the collection of all μ^*-measurable sets. Then $\mathcal{M}$ is a σ-algebra, and the restriction of μ^* to $\mathcal{M}$: $\mu = \mu^*|_{\mathcal{M}}, \mu(E) = \mu^*(E), E \in \mathcal{M}$ satisfies:

(a) $(O, \mathcal{M}, \mu)$ is a complete measure space.

(b) If $E \subset O$ and $\mu^*(E) = 0$, then $E \in \mathcal{M}$ and thus $\mu(E) = 0$.

From now on, we may also call a μ^*-measurable set a μ-measurable set. If μ^* is a metric outer measure, then

(c) $\mathcal{B}(O) \subset \mathcal{M}$, i.e. $\mathcal{M}$ contains all Borel sets of O and thus $(O, \mathcal{B}(O), \mu)$ where μ is implicitly further restricted to $\mathcal{B}(O)$ is a Borel measure space.

Theorem T.1.2. *Rademacher theorem*
Let $m, n \in \{1, 2, \ldots\}$, $E \subset \mathbb{R}^n$ be an open set, $f : E \to \mathbb{R}^m$ be a Lipschitz function, then f is differentiable $\mathcal{L}_n$ a.e. and the gradient ∇f is a measurable function.

Theorem T.1.3. *Whitney extension theorem*
Let $E \subset \mathbb{R}^m$, $m \in \{1, 2, \ldots\}$, E a closed set, $f : E \to \mathbb{R}$, $g : E \to \mathbb{R}^m$ be continuous functions. Denote

$$R(x, a) = \frac{f(x) - f(a) - g(a)(x - a)}{|x - a|}, \; x, a \in E, x \neq a.$$

If for all compact set $C \in E$,

$$\sup\{|R(x, a)|\, |0 < |x - a| \leqslant \delta, x, a \in C\} \to 0,$$

as $\delta \downarrow 0$. Then there exists a C^1 function $\overline{f} : \mathbb{R}^m \to \mathbb{R}$ such that

$$\overline{f}|_E = f, \quad \overline{f}'|_E = f'.$$

Theorem T.1.4. *Lusin theroem*
Let μ be a Borel regular measure over a metric space X, $m \in \{1, 2, \ldots\}$, $f : X \to \mathbb{R}^m$ be a μ

measurable function, $E \subset X$ *be a* μ *measurable set,* $\mu(E) < \infty$*. Then for arbitrary* $\epsilon > 0$*, there exists a compact set* $C \subset E$ *such that* $\mu(E \backslash C) < \epsilon$ *and* $f|_C$ *is continuous.*

Theorem T.1.5. *Egoroff theorem*
Let μ *be a Borel regular measure over a metric space* X*,* $m \in \{1, 2, \ldots\}$*,* $\{f_n\}$ *be a sequence of* μ *measurable functions* $f_n : X \to \mathbb{R}^m$*,* $f : X \to \mathbb{R}^m$ *be a* μ *measurable function. If*

$$f_n(x) \to f(x), \ \mu \text{ a.e. } x \in E,$$

where $E \subset X$ *is* μ *measurable,* $\mu(E) < \infty$*. Then for arbitrary* $\epsilon > 0$*, there exists a* μ *measurable set* $S \subset E$ *such that* $\mu(E \backslash S) < \epsilon$ *and*

$$f_n \to f, \ \text{uniformly on } S.$$

Definition T.1.1. (Vitali cover) Let $E \subset \mathbb{R}^m$, $m \in \{1, 2, \ldots\}$. If $\mathcal{V}$ is a collection of closed balls or closed cubes in $\mathbb{R}^m$ such that for all $x \in E$ and arbitrary $\epsilon > 0$, there exists $B \in \mathcal{V}$ such that $x \in B$ and $\mathrm{diam} B < \epsilon$, then $\mathcal{V}$ is called a Vitali cover of E.

Theorem T.1.6. *Vitali covering theorem*
Let $E \subset \mathbb{R}^m$*,* $\mathcal{V}$ *is a Vitali cover of* E*. Then, there exists an at most countable disjoint subset* $\{B_j\} \subset \mathcal{V}$*, such that*

$$\mathcal{L}_m^* \left(E \backslash \bigcup_j B_j \right) = 0.$$

Theorem T.1.7. *Isodiametric inequality*
For all set $E \subset \mathbb{R}^m$*,* $m \in \{1, 2, \ldots\}$*,*

$$\mathcal{L}^*(E) \leqslant \alpha_m \left(\frac{\mathrm{diam}\ E}{2} \right)^m.$$

Theorem T.1.8. *Coincidence between Spherical Hausdorff and Hausdorff outer measures*
For all set $E \subset \mathbb{R}^m$*,* $m \in \{1, 2, \ldots\}$*,*

$$\mathcal{H}_m^{S*} = \mathcal{H}_m^*,$$

where $\mathcal{H}_m^{S*}$*, the spherical Hausdorff outer measure is defined as*

$$\mathcal{H}_m^{S*} = \lim_{\delta \downarrow 0} \inf \left\{ \sum_{j \geqslant 1} \alpha_m \left(\frac{\mathrm{diam}\ B_j}{2} \right)^m : E \subset \bigcup_{j \geqslant 1} B_j, \mathrm{diam}\ B_j \leqslant \delta, B_j \text{ is a closed ball} \right\}.$$

Definition T.1.2 (Clarke Jacobian). Let $F : \Omega \to \mathbb{R}^m$ be a locally Lipschitz function, where $\Omega \subset \mathbb{R}^n$ is an open set. The Clarke Jacobian of F at $x \in \Omega$, denoted as $J_c F(x)$, is

$$J_c F(x) \coloneqq \operatorname{conv}\left\{ \lim_{k\to\infty} JF(x_k) : x_k \to x, JF(x_k) \text{ exists} \right\}.$$

Theorem T.1.9. *Nonsmooth implicit function theorem*
Let $F : \Omega \to \mathbb{R}^m$ be a locally Lipschitz (C^1) function, where $\Omega \subset \mathbb{R}^{n+m}$ is an open set. Assume that (x_0, y_0) is such that

1. *$F(x_0, y_0) = 0$.*
2. *$J_{c,y}F(x_0, y_0)$ is full rank in the sense that all matrices in $J_{c,y}F(x_0, y_0)$ is full rank, where $J_{c,y}F(x_0, y_0)$ consists of all $m \times m$ component matrices in $J_c F(x_0, y_0)$ written as $[A_{m\times n}, B_{m\times m}]$.*

Then there exists an $(n+m)$-dimensional interval $I = I_x^n \times I_y^m \subset \Omega$, where for some positive vectors α and β,

$$I_x = \{x \in \mathbb{R}^n : |x - x_0| < \alpha\}, \quad I_y = \{y \in \mathbb{R}^m : |y - y_0| < \beta\},$$

where $|x - x_0| < \alpha$ means that $|x_i - x_{0,i}| < \alpha_i$ for $\alpha = (\alpha_1, \ldots, \alpha_n)$, and a Lipschitz ($C^1$) function $\xi : I_x^n \to I_y^m$ such that for all $(x, y) \in I_x^n \times I_y^m$,

$$F(x, y) = 0 \Leftrightarrow y = \xi(x).$$

Theorem T.1.10. *Morse-Sard theorem*
Let $f \in C^r(U, \mathbb{R}^m)$, where $U \in \mathbb{R}^n$ is an open set, $r > \max(n-m, 0)$, then the set of critical values of f of zero Lebesgue measure and is meager.

Theorem T.1.11. *Partition of unity*
Let $E_1, \ldots, E_k$ be open sets in $\mathbb{R}^n$ and K a compact subset of $\bigcup_j K_j$. Then on can find $\phi_j \in C_0^\infty(E_j)$ so that $\phi_j \geqslant 0$ and $\sum_1^k \phi_j \leqslant 1$ with equality in a neighborhood of K.

Definition T.1.3. (Continuous convergence) Let $\mathcal{X}, \mathcal{Y}$ be two metric spaces, and $\{f_n\}$ be a sequence of mappings $f_n : \mathcal{X} \to \mathcal{Y}$. Given $f : \mathcal{X} \to \mathcal{Y}$ then f_n convergence to f continuously if $f_n(x_n) \to f(x)$ whenever $\{x_n\} \subset \mathcal{X}, x \in \mathcal{X}, x_n \to x$.

Lemma T.1.12. *Equivalence between uniform convergence and continuous convergence*
Let $\mathcal{X}$ and $\mathcal{Y}$ be two metric spaces, and $\{f_n\}, f$ be mappings from $\mathcal{X}$ to $\mathcal{Y}$.

1. *If $\mathcal{X}$ is compact and f is continuous then $f_n \to f$ continuously if and only if $f_n \to f$ uniformly in $\mathcal{X}$.*

2. If $f_n \to f$ continuously then f is continuous.

As a consequence, if $\mathcal{X}$ is compact then $f_n \to f$ continuously if and only if $f_n \to f$ uniformly in $\mathcal{X}$ and f is continuous.

PROOF OF LEMMA T.1.12.
1. If part: Let d be the metric on $\mathcal{Y}$. For $f_n \to f$ uniformly and $x_n \to x$ we have

$$\begin{aligned} d\left(f_n\left(x_n\right), f\left(x\right)\right) &\leqslant d\left(f_n\left(x_n\right), f\left(x_n\right)\right) + d\left(f\left(x_n\right), f\left(x\right)\right) \\ &\leqslant \sup_{x \in X} d\left(f_n\left(x_n\right), f\left(x_n\right)\right) + d\left(f\left(x_n\right), f\left(x\right)\right). \end{aligned}$$

As $n \to \infty$, the first term goes to 0 by uniform convergence and the second term goes to 0 by continuity. *Only if part:* Prove by contradiction. If $f_n \to f$ continuously but not uniformly. Then there is a subsequence $\{n_k\}$ and $\epsilon > 0$ such that for all n_k

$$\sup_{x \in X} d\left(f_{n_k}\left(x\right), f\left(x\right)\right) > 2\epsilon.$$

By the definition of sup there is a sequence $\{x_k\}$ such that

$$d\left(f_{n_k}\left(x_k\right), f\left(x_k\right)\right) > \epsilon. \tag{23}$$

Since $\mathcal{X}$ is a compact metric space there is a convergent subsequence $\{x_{k'}\}$ of $\{x_k\}$ with $x_{k'} \to x_0$. Continuous convergence and continuity of f require

$$d\left(f_{n_{k'}}\left(x_{k'}\right), f\left(x_{k'}\right)\right) \leqslant d\left(f_{n_{k'}}\left(x_{k'}\right), f\left(x_0\right)\right) + d\left(f\left(x_0\right), f\left(x_{k'}\right)\right) \to 0$$

which violates (23).
2. Let $d_{\mathcal{X}}$ be the metric on $\mathcal{X}$. Suppose $\{x_n\}$ is an arbitrary sequence that converges to x. For all k, consider sequence $\{f_m\left(x_{k,m}\right)\}$, where $d_{\mathcal{X}}\left(x_{k,m}, x_k\right) < \frac{1}{m}$. By continuous convergence,

$$\lim_{m \to \infty} f_m\left(x_{k,m}\right) = f\left(x_k\right).$$

Therefore there exist $n_k \in \mathbb{N}$, such that for all $n > n_k$,

$$d\left(f_n\left(x_{k,n}\right), f\left(x_k\right)\right) < \frac{1}{k}.$$

Next consider a sequence $\{y_n\}$

$$\{y_n\} = \left\{ \underbrace{x_{1,n_1}}_{n_1-terms}, \underbrace{x_{2,n_1+n_2}}_{n_2-terms}, \ldots, \underbrace{x_{k,\sum_{j=1}^{k} n_j}}_{n_k-terms}, \ldots \right\}.$$

Since $y_n \to x$ as $n \to \infty$ by continuous convergence $f_n(y_n) \to f(x)$. Write $\sum_{j=1}^k n_j$ as $\sum^k n_j$. Note that $y_{\sum^k n_j} = x_{k,\sum^k n_j}$, $f_{\sum^k n_j}\left(y_{\sum^k n_j}\right) = f_{\sum^k n_j}\left(x_{k,\sum^k n_j}\right)$, we have

$$\lim_{k\to\infty} f_{\sum^k n_j}\left(x_{k,\sum^k n_j}\right) = f(x).$$

Then by

$$\begin{aligned} d\left(f\left(x_k\right), f\left(x\right)\right) &\leqslant d\left(f\left(x_k\right), f_{\sum^k n_j}\left(x_{k,\sum^k n_j}\right)\right) + d\left(f_{\sum^k n_j}\left(x_{k,\sum^k n_j}\right), f\left(x\right)\right) \\ &\leqslant \frac{1}{k} + o\left(1\right), \end{aligned}$$

f is continuous. ■

For the Rademacher theorem T.1.2, see Theorem 2.10.43 in Federer (1969). For the Whitney extension theorem T.1.3, see Theorem 6.10 in Evans and Garzepy (2015) or Theorem 3.1.14 in Federer (1969). For the form of the Lusin theorem T.1.4 and the Egoroff theorem T.1.5 used here, see Theorem 2.3.5 and Theorem 2.3.7 in Federer (1969), respectively. For Vitali covering theorem T.1.6, see Theorem 1.10 in Falconer (1986). For the isodiametric inequality T.1.7, see Theorem 2.4 in Evans and Garzepy (2015) or Corollary 2.10.33 of Federer (1969). For the relationship between the spherical Hausdorff outer measure and the Hausdorff outer measure, see 2.10.6 in Federer (1969). For the nonsmooth implicit function theorem T.1.9, see Clarke (1976) and Theorem 11 of Hiriart-Urruty (1979). For more information about the Morse-Sard theorem T.1.10, we refer to Figalli (2008) and its renowned infinite dimensional version Sard-Smale theorem from Smale (1965). The form of the partition of unity theorem T.1.11 is taken from Theorem 2.2 in Grigoryan (2009). See also Theorem 3.5 of Grigoryan (2009) for a manifold version. Most of these results have been rewritten to comply with the convention and styles used in Evans and Garzepy (2015) and Lu (2019). The proof of part 1 of Lemma T.1.12 comes from Resnick (1987), we remove the separable space requirement in the premise. Actually, if $\mathcal{X}$ is a compact metric space, then $\mathcal{X}$ is separable, a countable dense subset of $\mathcal{X}$ can be constructed by the totally boundedness of $\mathcal{X}$.

T.2 Geometric interpretation of Hausdorff measure

In the following proposition, we explain intuitively how Hausdorff measure generalizes the usual conceptions of area and volume.

Proposition T.2.1. *Let $k \in \{1, 2, \ldots, n\}$, matrix $A \in \mathbb{R}^{n\times k}$, then*

(i) if $\operatorname{rank}(A) < k$, that is $\det\left(A^T A\right) = 0$, then $\mathcal{H}_k\left(A\left(E\right)\right) = 0$ for all $E \subset \mathbb{R}^k$.

(ii) *if* $\operatorname{rank}(A) = k$, *that is* $\det\left(A^T A\right) > 0$, *then for all* $E \subset \mathbb{R}^k$,

$$E \text{ is Lebesgue measurable} \Leftrightarrow A(E) \text{ is } \mathcal{H}_k \text{ measurable.}$$

When either of the sides holds, $\mathcal{H}_k(A(E)) = \sqrt{\det(A^T A)}\mathcal{L}_k(E)$.

Proposition T.2.2. *Let* $k \in \{1, 2, \ldots, n\}$, *matrix* $A \in \mathbb{R}^{k \times n}$, *then*

(i) *if* $\operatorname{rank}(A) < k$, *that is* $JA = \sqrt{\det(AA^T)} = 0$, *then* $\mathcal{H}_{n-k}\left(E \cap A^{-1}(y)\right) = 0$ *for* $\mathcal{L}_k$ *a.e.* $y \in \mathbb{R}^k$ *for all* $E \subset \mathbb{R}^n$.

(ii) *if* $\operatorname{rank}(A) = k$, *that is* $\det\left(AA^T\right) > 0$, *then for all measurable* $E \subset \mathbb{R}^n$, $y \in \mathbb{R}^k \mapsto \mathcal{H}_{n-k}\left(E \cap A^{-1}(y)\right)$ *is* $\mathcal{L}_k$ *measurable, and*

$$JA \cdot \mathcal{L}_n(E) = \int_E JA(x)\, d\mathcal{L}_n x = \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(E \cap A^{-1}(y)\right) d\mathcal{L}_k y.$$

Proposition T.2.1 provides the geometric meaning of the Hausdorff measure. Consider the set $A\left([0,1]^k\right), A \in \mathbb{R}^{n \times k}, k \in \{0, 1, \ldots, n\}$, denote the column vectors of A as $a_1, a_2, \ldots, a_k$. Let $e = (e_1, e_2, \ldots, e_k)^T \in E \subset \mathbb{R}^k$, then $Ae = \sum_{i=1}^k e_i a_i$, therefore

$$A\left([0,1]^k\right) = \left\{\sum_{i=1}^k e_i a_i | e = (e_1, e_2, \ldots, e_k)^T \in [0,1]^k\right\}$$

is the parallelepiped spanned by vectors $a_1, a_2, \ldots, a_k$. By Proposition T.2.1,

$$\mathcal{H}_k\left(A\left([0,1]^k\right)\right) = \sqrt{\det(A^T A)}.$$

We point out that $\sqrt{\det(A^T A)}$ is the k-dimensional volume (in n-dimensional space) of the parallelepiped spanned by the column vectors of matrix A. To see this, first consider the $k = n$ scenario. In this case $\sqrt{\det(A^T A)} = |\det(A)|$. Denote the volume of the parallelepiped spanned by A as $\operatorname{vol}(A)$, or more transparently as $\operatorname{Vol}(a_1, a_2, \ldots, a_k)$ where $\{a_i\}_{i=1}^k$ are the columns of A. The volume of a parallelepiped is the "area" of its base, times its height. A base is the parallelepiped determined by arbitrarily chosen $k-1$ vectors from $\{a_i\}_{i=1}^k$, and the height corresponding to this base is the perpendicular distance of the remaining vector from the base. Denote the remaining vector as a_i. If a_i is scaled by a factor of c, then the perpendicular distance of a_i from the base and thus the volume will be scaled by a factor of $|c|$. If a_i is translated to $a_i' = a_i + \omega a_j, i \neq j$, since a_j is parallel to the base, the height and thus the volume will not change[4], i.e.

$$\operatorname{Vol}(a_1, \ldots, ca_i + \omega a_j, \ldots, a_k) = |c|\operatorname{Vol}(a_1, \ldots, a_i, \ldots, a_k). \tag{24}$$

[4]See https://textbooks.math.gatech.edu/ila/determinants-volumes.html for a visualization.

Swapping two columns of A just reorders the vectors $\{a_i\}_{i=1}^k$ and will not change the volume,

$$\mathrm{Vol}\left(a_1, \ldots, a_i, \ldots, a_j, \ldots, a_k\right) = \mathrm{Vol}\left(a_1, \ldots, a_j, \ldots, a_i, \ldots, a_k\right). \tag{25}$$

Since $|\det(A)|$ can also be characterized by the properties (24) and (25) and $|\det(I_k)| = \mathrm{Vol}(I_k) = 1$ where I_k is the $k \times k$ identity matrix, we have $\sqrt{\det(A^T A)} = |\det(A)| = \mathrm{Vol}(A)$.

When $k < n$, using the Singular Value Decomposition $A = U\Sigma V$ where U, V are orthogonal matrixes and Σ a rectangular diagonal matrix with non-negative real diagonal:

$$\sqrt{\det(A^T A)} = \sqrt{\det\left((U\Sigma V)^T U\Sigma V\right)} = \sqrt{\det(\Sigma^T \Sigma)}$$

where $\Sigma^T\Sigma$ is a $k \times k$ diagonal matrix. We claim that the k-dimensional volume (in n-dimensional space) of A is also $\sqrt{\det(\Sigma^T\Sigma)}$. In particular, since orthogonal transformations preserves inner products and thus lengths (norm) and angles,

$$\mathrm{Vol}(A) = \mathrm{Vol}(U\Sigma V) = \mathrm{Vol}(\Sigma V).$$

Note that, if we choose a orientation for a base and allow for signed height and volume, then we have

$$\mathrm{Vol}_s\left(a_1, \ldots, a_i + \Delta a_i, \ldots, a_k\right) = \mathrm{Vol}_s\left(a_1, \ldots, a_i, \ldots, a_k\right) + \mathrm{Vol}_s\left(a_1, \ldots, \Delta a_i, \ldots, a_k\right)$$

and $\mathrm{Vol}(A) = |\mathrm{Vol}_s(A)|$[5]. Signed volume also satisfies homogeneity similar to (24), with the scale factor changing from $|c|$ to c. Then we have

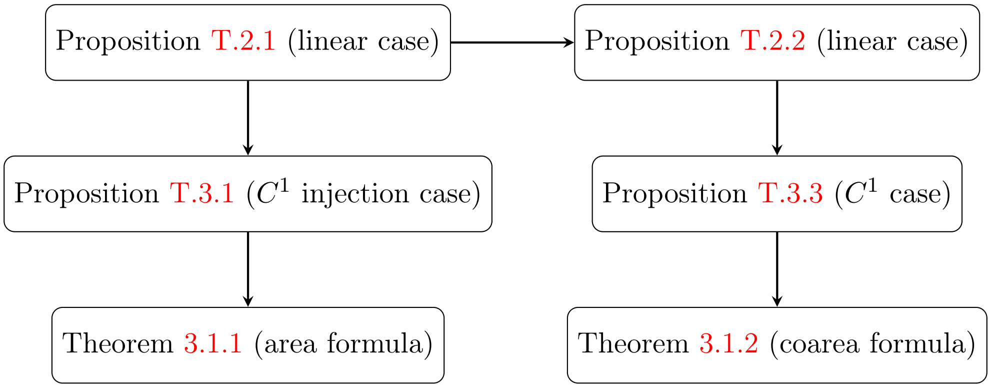
$$\begin{aligned}\mathrm{Vol}_s\left(a_1, a_2, \ldots, a_k\right) &= \mathrm{Vol}_s\left(\sum_{i_1=1}^{n} a_{i_1,1} e_{i_1}, \sum_{i_2=1}^{n} a_{i_2,2} e_{i_2}, \ldots, \sum_{i_k=1}^{n} a_{i_k,k} e_{i_k}\right)\\ &= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_k}^{n}\prod_{j=1}^{k} a_{i_j,j}\mathrm{Vol}_s\left(e_{i_1}, e_{i_2}, \ldots, e_{i_k}\right),\end{aligned}$$

where each e_{i_j} is a $n \times 1$ vector with 1 in the i_jth position and zeros otherwise. When $\{i_j\}_{j=1}^k$ contains repeated values, $\mathrm{rank}\left(e_{i_1}, e_{i_2}, \ldots, e_{i_k}\right) < k$ and $\mathrm{Vol}_s\left(e_{i_1}, e_{i_2}, \ldots, e_{i_k}\right) = 0$. Therefore, scaling a row of $n \times k$ matrix by a factor of d also scales the volume by the factor of $|d|$, implying that $\mathrm{Vol}(\Sigma V) = \prod_{i=1}^k \sigma_i \mathrm{Vol}(V)$, where σ_i are the diagonal elements of Σ. Intuitively, $\mathrm{Vol}\left([0,1]^k\right) = 1$ since volume is base times height. Thus, the volume of $\mathrm{Vol}(V)$ is equal to 1.

For an matrix $A \in \mathbb{R}^{k \times n}$, consider the geometric meaning of JA. We already know that the k-dimensional volume of $A^T\left([0,1]^k\right)$ is $\sqrt{\det(AA^T)}$. For arbitrary $P \in \mathbb{R}^{n \times k}$, the k-dimensional

[5] We do not explicitly distinguish k-dimensional volume and n-dimensional volume in n-dimensional space, they are both denoted as Vol. The specific meaning is clear based on the context.

Figure 2: Ordering of proofs for area and coarea formulas



volume of $AP\left([0,1]^k\right)$ is equal to the k-dimensional volume of the $AP^{\perp}\left([0,1]^k\right)$, where $P^{\perp}$ is the orthogonal projection of P onto the orthogonal complement of the null space of A (which, by definition, is the span of the columns of A^T). Note that $\det\left(\left(P^{\perp}\right)^T P^{\perp}\right) \leqslant \det\left(P^T P\right)$. Therefore,

$$JA = \sup_P \frac{\mathrm{Vol}\,(AP)}{\mathrm{Vol}\,(P)} = \sup_P \frac{\mathcal{H}_k\left(AP\left([0,1]^k\right)\right)}{\mathcal{H}_k\left(P\left([0,1]^k\right)\right)},$$

where supremum is taken over all k-dimensional nondegenerate parallelepiped P.

T.3 Proofs of area and coarea formulas

Figure 2 illustrates the ordering of the proofs for area and coarea formulas. Besides Federer (1969) and Evans and Garzepy (2015), we also borrow lots of material from Jerrard (2013) and Lu (2019), especially from the latter. Corollaries T.3.2 and T.3.4 are used to illustrate the applicability of the area and coarea formulas 3.1.1 and 3.1.2.

Proof Outline of Proposition T.2.1.

step 1 The Hausdorff outer measure $\mathcal{H}_k^*$ is invariant under orthogonal transformations, i.e. if A is a orthogonal transformations then

$$\mathcal{H}_k^*\left(A\left(E\right)\right) = \mathcal{H}_k^*\left(E\right), \forall E \subset \mathbb{R}^k$$

step 2 Let $A = \left(I_k, 0\right)^T$, prove that

$$\mathcal{H}_k^*\left(A\left(E\right)\right) = \mathcal{H}_k^*\left(E \times \{0\}\right) = \mathcal{H}_k^*\left(E\right) = \mathcal{L}_k^*\left(E\right)$$

where $\mathcal{L}_k^*$ is the k-dimensional Lebesgue outer measure.

step 3 There exists orthogonal $T \in \mathbb{R}^{n\times n}$,

$$\mathcal{H}_k^*(A(E)) = \mathcal{H}_k^*(TA(E)) = \mathcal{H}_k^*((A_1(E)) \times \{0\}) = \mathcal{L}_k^*(A_1(E)),$$

where $A_1 \in \mathbb{R}^{k\times k}$, $A \in \mathbb{R}^{n\times k}$, and $TA = \left(A_1^T, 0\right)^T$ **step** 4 Proof of

$$\mathcal{L}_k^*(A_1(E)) = |\det(A_1)|\mathcal{L}_k^*(E) = \sqrt{\det(A^TA)}\mathcal{L}_k^*(E).$$

step 5 For all $E \subset \mathbb{R}^k$, E is Lebesgue measurable $\Leftrightarrow A(E)$ is $\mathcal{H}_k$ measurable. ■

Remark T.3.1. We provide two supplementary details to assist with understanding the proof outlines of Proposition T.2.1 and the following results.

1. Hausdorff outer measure $\mathcal{H}_k^*$ has the invariance under the orthogonal transformations. This is a direct result of the fact that Hasudroff outer measure keeps distance inequality, i.e. if $k, l, m, n \in \mathbb{Z}^+$, $E \subset \mathbb{R}^l$, map $f : E \to \mathbb{R}^n$ and map $g : E \to \mathbb{R}^m$ satisfies

$$|f(x) - f(y)| \leqslant C|g(x) - g(y)| \quad \forall x, y \in E,$$

then $\mathcal{H}_k^*(f(E)) \leqslant C^k\mathcal{H}_k^*(g(E))$. This property will also be used in the following discussion.

2. A cube is a subset of $\mathbb{R}^k$ in the form of

$$\prod_{i=1}^k(a_i, b_i), \quad \prod_{i=1}^k(a_i, b_i], \quad \prod_{i=1}^k[a_i, b_i), \quad \prod_{i=1}^k(a_i, b_i) \quad a_i < b_i, \forall i \in \{1, 2, \ldots, k\}.$$

Let $E \in \mathbb{R}^n$ be a nonempty open set, then there exists a sequence of disjoint left open and right closed cubes $\{Q_k\}_{k=1}^\infty$, such that

$$E = \bigcup_{k=1}^\infty Q_k = \bigcup_{k=1}^\infty \bar{Q}_k.$$

Proposition T.3.1. *Let $E \subset \mathbb{R}^k$ be an open set, $k \in \{1, 2, \ldots, n\}$, $\psi : E \to \mathbb{R}^n$ be a C^1 injection and* $\det\left(\psi'(x)^T\psi'(x)\right) > 0$ *for all $x \in E$, then for all $D \subset \psi(E)$,*

$$D \text{ is } \mathcal{H}_k \text{ measurable} \Leftrightarrow \psi^{-1}(D) \text{ is Lebesgue measurable.}$$

When D is $\mathcal{H}_k$ measurable,

$$\mathcal{H}_k(D) = \int_{\psi^{-1}(D)} \sqrt{\det\left(\psi'(x)^T\psi'(x)\right)}d\mathcal{L}_kx.$$

Proposition T.3.1 plays the most essential role in a proof of more general area formula Theorem 3.1.1. Proposition T.3.1 can also be powerful when used alone.

Proof Outline of Proposition T.3.1.

step 1 Estimates of $|\psi(x) - \psi(y)|$: Let $k \in \{1, 2, \ldots, n\}$, $E \subset \mathbb{R}^k$ be an open set, $\psi : E \to \mathbb{R}^n$ be a C^1 injection such that $\det\left(\psi'(x)^T \psi'(x)\right) > 0$ for all $x \in E$, then:

1. if $K \subset E$ is convex and compact, then there exists $0 < c < C$ such that

$$c|x - y| \leqslant |\psi(x) - \psi(y)| \leqslant C|x - y| \tag{26}$$

   for all $x, y \in K$.

2. For arbitrary $x_0 \in E$ and for all $0 < \epsilon < 1$, there exists $\delta > 0$, such that open ball $B(x_0, \delta) \subset E$ and

$$(1 - \epsilon)|\psi'(x_0)(x - y)| \leqslant |\psi(x) - \psi(y)| \leqslant (1 + \epsilon)|\psi'(x_0)(x - y)| \tag{27}$$

   for all $x, y \in B(x_0, \delta)$.

step 2 Prove that, for all $D \subset \psi(E)$,

$$D \text{ is } \mathcal{H}_k \text{ measurable} \Leftrightarrow \psi^{-1}(D) \text{ is Lebesgue measurable},$$

by estimates (26), the fact that Hausdorff outer measure keeps distance inequality, and the fact that if D is a $\mathcal{H}_k$ measurable set with $\mathcal{H}_k(D) < \infty$, then there exist a Borel set P and a $\mathcal{H}_k$ zero measure set Z such that $D = P \cup Z$[6].

step 3 By (27) and Proposition T.2.1, closed cube $Q \subset E$ satisfies: for all $x_0 \in E$, for all $0 < \epsilon < 1$, there exists $\delta > 0$, such that, if $\operatorname{diam}(Q) < \delta$ and $x_0 \in Q$ then

$$\begin{aligned}(1 - \epsilon)^k \sqrt{\det\left(\psi'(x_0)^T \psi'(x_0)\right)} \mathcal{L}_k(Q) &\leqslant \mathcal{H}_k(\psi(Q)) \\ &\leqslant (1 + \epsilon)^k \sqrt{\det\left(\psi'(x_0)^T \psi'(x_0)\right)} \mathcal{L}_k(Q)\end{aligned} \tag{28}$$

step 4 Prove that, for all closed cube $Q \in E$,

$$\mathcal{H}_k(\psi(Q)) = \int_Q \sqrt{\det\left(\psi'(x)^T \psi'(x)\right)} d\mathcal{L}_k x$$

[6] This Borel set, zero measure set construction of $\mathcal{H}_k$ measurable set is also a common useful result. Actually, the complete result states that there exist Borel sets P_1, P_2 and $\mathcal{H}_k$ zero measure sets Z_1, Z_2, such that $D = P_1 \cup Z_1 = P_2 \backslash Z_2$.

by (28).

step 5 Prove that for all bounded open set E_b such that $\overline{E_b} \subset E$, if $O \subset E_b$ is $\mathcal{L}_k$ measurable then

$$\mathcal{H}_k\left(\psi\left(O\right)\right) = \int_O \sqrt{\det\left(\psi'\left(x\right)^T \psi'\left(x\right)\right)} d\mathcal{L}_k x.$$

Conclude using the fact that any open set $E \subset \mathbb{R}^k$ can be decomposed to a countable disjoint union of bounded cube. ■

Proof of Theorem 3.1.1.

A classical proof based on Proposition T.3.1 can be separated into three fundamental parts.

part 1 In case that ψ is not necessarily bijective, while still requiring that $J\psi\left(x\right) > 0$ for all $x \in E$, by the implicit function theorem, for all $x \in E$ there exist a neighborhood U such that ψ is bijective in U. Take a Vitali cover $\mathcal{V}$ of E such that ψ is bijective in every closed ball $B \in \mathcal{V}$. Then by the Vitali covering theorem, there exists an at most countable disjoint subset $\{B_j\} \subset \mathcal{V}$, such that $\mathcal{L}_k^*\left(E \backslash \bigcup_j B_j\right) = 0$. From the definition of $\mathcal{H}_0$,

$$\sum_j 1\left(y \in \psi\left(S \cap B_j\right)\right) = \mathcal{H}_0\left(\bigcup_j \left(S \cap B_j\right) \cap \psi^{-1}\left(y\right)\right).$$

By the property of Lebesgue integral and Proposition T.3.1,

$$\int_S J\psi\left(x\right) d\mathcal{L}_k x = \sum_j \int_{\psi(S)} 1\left(y \in \psi\left(S \cap B_j\right)\right) d\mathcal{H}_k y = \int_{\psi(S)} \mathcal{H}_0\left(\bigcup_j \left(S \cap B_j\right) \cap \psi^{-1}\left(y\right)\right) d\mathcal{H}_k y.$$

Then note that

$$\begin{aligned}
&\int_{\psi(S)} \mathcal{H}_0\left(S \cap \psi^{-1}\left(y\right)\right) d\mathcal{H}_k y \\
&= \int_{\psi(S)} \mathcal{H}_0\left(\bigcup_j \left(S \cap B_j\right) \cap \psi^{-1}\left(y\right)\right) d\mathcal{H}_k y + \int_{\psi(S)\backslash\psi\left(S \backslash \bigcup_j B_j\right)} \mathcal{H}_0\left(\left(S \backslash \bigcup_j B_j\right) \cap \psi^{-1}\left(y\right)\right) d\mathcal{H}_k y \\
&\qquad + \int_{\psi\left(S \backslash \bigcup_j B_j\right)} \mathcal{H}_0\left(\left(S \backslash \bigcup_j B_j\right) \cap \psi^{-1}\left(y\right)\right) d\mathcal{H}_k y \\
&= \int_{\psi(S)} \mathcal{H}_0\left(\bigcup_j \left(S \cap B_j\right) \cap \psi^{-1}\left(y\right)\right) d\mathcal{H}_k y + 0 + 0.
\end{aligned}$$

When $J\psi\left(x\right) = 0$, let $\epsilon > 0$, define $\psi_\epsilon : E \to \mathbb{R}^{k+n}$ as

$$x \mapsto \left(\epsilon x, \psi\left(x\right)\right),$$

$Crit\left(\psi\right)=\{x\in E|J\psi\left(x\right)=0\}$. Note that $J\psi_{\epsilon}\left(x\right)>0$ for all $x\in E$,

$$\begin{aligned}\int_{Crit(\psi)}J\psi_{\epsilon}\left(x\right)d\mathcal{L}_{k}x&=\int_{\psi_{\epsilon}(Crit(\psi))}\mathcal{H}_{0}\left(Crit(\psi)\cap\psi_{\epsilon}^{-1}\left(y\right)\right)d\mathcal{H}_{k}y\\&\geqslant\int_{\psi_{\epsilon}(Crit(\psi))}d\mathcal{H}_{k}y=\mathcal{H}_{k}\left(\psi_{\epsilon}\left(Crit\left(\psi\right)\right)\right).\end{aligned}$$

By the fact that Hausdorff outer measure keeps distance inequality,

$$\mathcal{H}_{k}\left(\psi\left(Crit\left(\psi\right)\right)\right)\leqslant\mathcal{H}_{k}\left(\psi_{\epsilon}\left(Crit\left(\psi\right)\right)\right),$$

since the coordinate projection from $\psi_{\epsilon}\left(Crit\left(\psi\right)\right)$ to $\psi\left(Crit\left(\psi\right)\right)$ satisfies a distance inequality with $C=1$. Therefore,

$$\mathcal{H}_{k}\left(\psi\left(Crit\left(\psi\right)\right)\right)\leqslant\int_{Crit(\psi)}J\psi_{\epsilon}\left(x\right)d\mathcal{L}_{k}x,$$

the right hand side converges to 0 as $\epsilon\downarrow 0$ if E is bounded. Then conclude that $\mathcal{H}_{k}\left(\psi\left(Crit\left(\psi\right)\right)\right)=0$ for (not necessarily bounded) open set $E\subset\mathbb{R}^{k}$ by the cube decomposition in the step 5 of the sketch of the proof of Proposition T.3.1. Now, one can conclude that (6) is true for C^{1} function ψ.

part 2 To verify (6) when ψ is a Lipschitz function but does not necessarily belong to C^{1}, a Lusin type approximation of ψ can be used. To continue, we use Rademacher theorem T.1.2, Whitney extension theorem T.1.3, Lusin theroem T.1.4 and Egoroff theorem T.1.5.

Assume first E is bounded, by Rademacher theorem T.1.2, ψ is differentiable $\mathcal{L}_{k}$ *a.e.* and the gradient $\psi'\leqslant\mathrm{Lip}\psi$ is measurable, where

$$\mathrm{Lip}\psi=\sup\left\{\frac{\left|\psi\left(x_{1}\right)-\psi\left(x_{2}\right)\right|}{\left|x_{1}-x_{2}\right|}|x_{1},x_{2}\in E,x_{1}\neq x_{2}\right\}.$$

Apply Lusin theorem to ψ', there exists a compact set $C\subset E$ such that $\mathcal{L}_{k}\left(E\backslash C\right)<\frac{1}{2}\epsilon$ and $\psi'|_{C}$ is continuous. Let

$$R\left(x,a\right)=\frac{\psi\left(x\right)-\psi\left(a\right)-\psi'\left(a\right)\left(x-a\right)}{\left|x-a\right|},\ x,a\in C,x\neq a,$$

since ψ is differentiable, for all $a\in C$,

$$R\left(a\right)=\sup\{\left|R\left(x,a\right)\right||0<\left|x-a\right|\leqslant\delta,x\in C\}\to 0,$$

as $\delta\downarrow 0$. Then by Egoroff theorem and regularity of Lebesgue measure, there exists a compact set $C'\subset C$ such that $\mathcal{L}_{k}\left(C\backslash C'\right)<\frac{1}{2}\epsilon$ and $R\left(a\right)$ converge to 0 uniformly on C'. Now, we can apply Whitney extension theorem to ψ and ψ' (actually, to each component function of ψ and its gradient), i.e. there exists a C^1 function $\overline{\psi}$ such that

$$\begin{aligned} \overline{\psi}|_{C'}(x) = \psi|_{C'}(x), \quad \overline{\psi}'|_{C'}(x) = \psi'|_{C'}(x), \\ \mathcal{L}_k\left(\left\{x|\overline{\psi}|_E(x) \neq \psi(x)\right\}\right) < \epsilon, \quad \mathcal{L}_k\left(\left\{x|\overline{\psi}'|_E(x) \neq \psi'(x)\right\}\right) < \epsilon. \end{aligned} \tag{29}$$

Now, one can conclude that (6) is true for Lipschitz function ψ and set $S \cap C'$.

part 3 The final step is to verify a Lusin property (N) for

$$\int_{\psi(S)} \mathcal{H}_0\left(S \cap \psi^{-1}(y)\right) d\mathcal{H}_k y.$$

Specifically, for arbitrary measurable $S \subset E$,

$$\int_{\psi(S)} \mathcal{H}_0\left(S \cap \psi^{-1}(y)\right) d\mathcal{H}_k y \leqslant (\mathrm{Lip}\psi)^n \mathcal{L}_k(S).$$

To see this, let

$$\mathcal{Q}_m = \left\{Q|Q = \prod_{i=1}^k (a_i, b_i], a_i = \frac{c_i}{m}, b_i = \frac{c_i+1}{m}, c_i \in \mathbb{Z}\right\},$$

and

$$g_m(y) = \sum_{Q \in \mathcal{Q}_m} 1\left(y \in \psi(S \cap Q)\right).$$

Since $\mathbb{R}^k = \bigcup_{Q \in \mathcal{Q}_m}$, and $g_m(y)$ is the number of cubes $Q \in \mathcal{Q}_m$ such that

$$\mathcal{H}_0\left(S \cap Q \cap \psi^{-1}(y)\right) > 0.$$

Therefore, for all $y \in \mathbb{R}^n$,

$$g_m(y) \uparrow \mathcal{H}_0\left(S \cap \psi^{-1}(y)\right),$$

as $m \to \infty$. Then by the monotone convergence theorem

$$\begin{aligned} \int_{\psi(S)} \mathcal{H}_0\left(S \cap \psi^{-1}(y)\right) d\mathcal{H}_k y &= \lim_{m\to\infty} \int_{\psi(S)} g_m(y)\, d\mathcal{H}_k y \\ &= \lim_{m\to\infty} \sum_{Q \in \mathcal{Q}_m} \mathcal{H}_k\left(\psi(S \cap Q)\right) \\ &\leqslant \lim_{m\to\infty} \sum_{Q \in \mathcal{Q}_m} (\mathrm{Lip}\psi)^k \mathcal{L}_k(S \cap Q) \\ &= (\mathrm{Lip}\psi)^k \mathcal{L}_k(S). \end{aligned}$$

Now note that,

$$\int_{S\cap C'} J\psi\left(x\right)d\mathcal{L}_k x \leqslant \int_{\psi(S)} \mathcal{H}_0\left(S\cap\psi^{-1}\left(y\right)\right)d\mathcal{H}_k y \leqslant \int_{S\cap C'} J\psi\left(x\right)d\mathcal{L}_k x + \left(\mathrm{Lip}\psi\right)^k \mathcal{L}_k\left(S\backslash C'\right),$$

where $\mathcal{L}_k\left(S\backslash C'\right) < \epsilon$. Note that $J\psi$ is bounded on E due to the Lipschitz continuity of ψ (By Rademacher theorem, $J\psi$ exists for $\mathcal{L}_k$ *a.e.* $x \in E$), i.e. there exists a constant M, such that

$$\int_{S\backslash C'} J\psi\left(x\right)d\mathcal{L}_k x \leqslant M\mathcal{L}_k\left(S\backslash C'\right).$$

Therefore,

$$\begin{aligned}\int_{S\cap C'} J\psi\left(x\right)d\mathcal{L}_k x &\leqslant \int_{S} J\psi\left(x\right)d\mathcal{L}_k x \\ &\leqslant \int_{S\cap C'} J\psi\left(x\right)d\mathcal{L}_k x + M\mathcal{L}_k\left(S\backslash C'\right),\end{aligned}$$

and (6) follows from the arbitrariness of ϵ. The case when open set $E\subset\mathbb{R}^k$ is unbounded follows from the cube decomposition. ■

Corollary T.3.2. *A Sard type lemma*
Let $k\in\{1,2,\ldots,n\}$, $E\subset\mathbb{R}^k$ be an open set, $\psi: E\to\mathbb{R}^n$ be a C^1 function and $Crit\left(\psi\right) = \{x\in E|J\psi\left(x\right)=0\}$, then $\mathcal{H}_k\left(\psi\left(Crit\left(\psi\right)\right)\right) = 0$.

Proof. Since ψ is a C^1 function, $J\psi: E\to\mathbb{R}$ is continuous. Therefore, $Crit\left(\psi\right) = \{x\in E|J\psi\left(x\right)\geqslant 0\}\backslash\{x\in E|J\psi\left(x\right)>0\}$ is $\mathcal{L}_k$ measurable. By (6),

$$\begin{aligned}\int_{Crit(\psi)} J\psi\left(x\right)d\mathcal{L}_k x &= \int_{\psi(Crit(\psi))} \mathcal{H}_0\left(Crit(\psi)\cap\psi^{-1}\left(y\right)\right)d\mathcal{H}_k y \\ &\geqslant \int_{\psi(Crit(\psi))} d\mathcal{H}_k y = \mathcal{H}_k\left(\psi\left(Crit\left(\psi\right)\right)\right).\end{aligned}$$

The integral in the left hand side of the first equality is 0. ■

Proof Outline of Proposition T.2.2.
step 1 To prove *(i)*, note that

$$\mathcal{L}_k^*\left(A\left(\mathbb{R}^n\right)\right) = \mathcal{L}_k^*\left(U\circ\Sigma\circ V\left(\mathbb{R}^n\right)\right) = \mathcal{L}_k^*\left(U\circ\Sigma\left(\mathbb{R}^n\right)\right) = \mathcal{L}_k^*\left(\Sigma\left(\mathbb{R}^n\right)\right),$$

by SVD and the fact that Lebesgue outer measure is invariant under orthogonal transformation.

Since $\operatorname{rank}(A) < k$, $\operatorname{rank}(\Sigma) < k$, therefore,

$$\mathcal{L}_k^*\left(\Sigma\left(\mathbb{R}^n\right)\right) = \mathcal{L}_k^*\left(\mathbb{R}^{\operatorname{rank}(\Sigma)}\right) = 0.$$

step 2 By SVD / PD, $A = WPV$, where $V \in \mathbb{R}^{n\times n}$ is orthogonal, $P : \mathbb{R}^n \to \mathbb{R}^k$ is the coordinate projection of the first k dimensions, and $W \in \mathbb{R}^{k\times k}$ is symmetric[7]. **step** 3 By the Fubini-Tonelli theorem, $y \in \mathbb{R}^k \mapsto \mathcal{L}_{n-k}\left(\left\{(x_1,\ldots,x_{n-k}) : x \in V(E) \cap (P)^{-1}(y)\right\}\right)$ is $\mathcal{L}_k$ measurable and

$$\mathcal{L}_n(E) = \mathcal{L}_n(V(E)) = \int_{\mathbb{R}^k} \mathcal{L}_{n-k}\left(\left\{(x_1,\ldots,x_{n-k}) : x \in V(E) \cap (P)^{-1}(y)\right\}\right) d\mathcal{L}_k y.$$

Then, note that

$$\begin{aligned}\mathcal{L}_{n-k}\left(\left\{(x_1,\ldots,x_{n-k}) : x \in V(E) \cap (P)^{-1}(y)\right\}\right) &= \mathcal{H}_{n-k}\left(V(E) \cap (P)^{-1}(y)\right)\\ &= \mathcal{H}_{n-k}\left(E \cap A^{-1} \circ W(y)\right)\end{aligned}$$

by $A^{-1} = V^{-1} \circ (P)^{-1} \circ W^{-1}$.

step 4 By Proposition T.2.1 and a standard approximation procedure,

$$JM \cdot \int_{\mathbb{R}^k} f(M(x))\, d\mathcal{L}_k x = \int_{\mathbb{R}^k} f(y)\, d\mathcal{L}_k y.$$

for all $M \in \mathbb{R}^{k\times k}$ invertible and f nonnegative $\mathcal{L}_k$ measurable. Therefore,

$$JW \cdot \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(E \cap A^{-1} \circ W(y)\right) d\mathcal{L}_k y = \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(E \cap A^{-1}(y)\right) d\mathcal{L}_k y.$$

■

Proposition T.3.3. *Let $E \subset \mathbb{R}^n$ be an open set, $k \in \{1,2,\ldots,n\}$, $\varphi : E \to \mathbb{R}^k$ be a C^1 function, then for all measurable S, $S \cap \varphi^{-1}(y)$ is $\mathcal{H}_{n-k}$ measurable for $\mathcal{H}_k$ a.e. y, $y \in \mathbb{R}^k \mapsto \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right)$ is $\mathcal{H}_k$ measurable, and*

$$\int_S J\varphi(x)\, d\mathcal{L}_n x = \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y.$$

Remark T.3.2.

1. We should note that for open set $E \subset \mathbb{R}^n$, $k \in \{1,2,\ldots,n\}$, $\varphi : E \to \mathbb{R}^k$ be at least continuous, $S \subset E$ measurable, then, $\varphi(S)$ **is not necessarily $\mathcal{H}_k$ measurable**. Actually,

[7] We do not distinguish between linear transformation and its matrix.

if $\varphi : E \subset \mathbb{R}^n \to \mathbb{R}^m, m, n \in \{1, 2, \ldots\}$, φ is continuous, then

$$\varphi \text{ map } \mathcal{L}_n \text{ measurable subset of } E \text{ to } \mathcal{L}_m \text{ measurable set in } \mathbb{R}^m \Leftrightarrow$$
$$\varphi \text{ map } \mathcal{L}_n \text{ zero measure subset of } E \text{ to } \mathcal{L}_m \text{ zero measure set in } \mathbb{R}^m.$$

Even if φ is more smooth than continuous, the right hand side of above relationship will not be automatically satisfied.

2. Although $S \cap \varphi^{-1}(y)$ may not be $\mathcal{H}_{n-k}$ measurable for all $y \in \mathbb{R}^k$, $E \cap \varphi^{-1}(y)$ is $\mathcal{H}_{n-k}$ measurable for all $y \in \mathbb{R}^k$, since E is open and $\varphi^{-1}(y)$ is a Borel set.

Proposition T.3.3 states one of the most essential idea of more general coarea formula Theorem 3.1.2. Besides, we should first verify that the integrand on the right hand side of (8) is well defined. A classical proof of Proposition T.3.3 can be separated into two fundamental parts.

Proof of Proposition T.3.3.
part 1 We start from verifying a Lusin property (N) for

$$\int_{\mathbb{R}^k} \mathcal{H}^*_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y.$$

First of all, the outer integral of f is defined as

$$\int^*_{\mathbb{R}^k} f(x)\, d\mathcal{H}_k x = \inf\left\{\int_{\mathbb{R}^k} g(x)\, d\mathcal{H}_k x | g \text{ is } \mathcal{H}_k \text{ measurable }, f \leqslant g \ a.e.\right\}.$$

For arbitrary measurable $S \subset E$, the required Lusin property (N) is provided by Eilenberg inequality which states that

$$\int^*_{\mathbb{R}^k} \mathcal{H}^*_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y \leqslant \frac{\alpha_{n-k}\alpha_k}{\alpha_n} (\mathrm{Lip}\varphi)^k \mathcal{H}_n(S) \tag{30}$$

holds. Besides, we can also show that $\mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right)$ is well defined and $\mathcal{H}_k$ measurable, and therefore the outer integral in (30) is actually redundant.

To verify (30), we use the isodiametric inequality T.1.7 and the coincidence between spherical Hausdorff outer measure and Hausdorff outer measure T.1.8. By the equivalence in Theorem T.1.8, for all $l > 0$, there exists an at most countable collection of closed balls $\{B^l_j\}$ such that $S \subset \bigcup_j B^l_j$, diam $B^l_j < \frac{1}{l}$ for all j, and

$$\sum_i \alpha_n \left(\frac{\text{diam } B^l_j}{2}\right)^n \leqslant \mathcal{H}_n(S) + \frac{1}{j}.$$

Define

$$g_j^l(y) = \alpha_{n-k} \left(\frac{\text{diam } B_j^l}{2} \right)^{n-k} 1\left(y \in \varphi\left(B_j^l\right) \right),$$

since B_j^l is a closed ball, g_j^l is $\mathcal{H}^k$ measurable. Note that $\{B_j^l\}$ covers $A \cap \varphi^{-1}(y)$ for all y,

$$\mathcal{H}^*_{n-k,\frac{1}{l}}\left(S \cap \varphi^{-1}(y)\right) \leqslant \sum_j g_j^l(y).$$

Then by the Fatou lemma, the isodiametric inequality, and the fact that Hausdorff outer measure keeps distance inequality,

$$\begin{aligned}
\int^*_{\mathbb{R}^k} \mathcal{H}^*_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y &= \int^*_{\mathbb{R}^k} \lim_{l\to\infty} \mathcal{H}^*_{n-k,\frac{1}{l}}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y \\
&\leqslant \int_{\mathbb{R}^k} \liminf_{l\to\infty} \sum_j g_j^l(y)\, d\mathcal{H}_k y \\
&\leqslant \liminf_{l\to\infty} \int_{\mathbb{R}^k} \sum_j g_j^l(y)\, d\mathcal{H}_k y \\
&\leqslant \liminf_{l\to\infty} \sum_j \int_{\mathbb{R}^k} g_j^l(y)\, d\mathcal{H}_k y \\
&\leqslant \liminf_{l\to\infty} \sum_j a_{n-k} \left(\frac{\text{diam } B_j^l}{2} \right)^{n-k} \mathcal{H}_k\left(\varphi\left(B_j^l\right)\right) \\
&= \liminf_{l\to\infty} \sum_j a_{n-k} \left(\frac{\text{diam } B_j^l}{2} \right)^{n-k} \mathcal{L}_k\left(\varphi\left(B_j^l\right)\right) \\
&\leqslant \liminf_{l\to\infty} \sum_j a_{n-k} \left(\frac{\text{diam } B_j^l}{2} \right)^{n-k} \alpha_k \left(\frac{\text{diam } \varphi\left(B_j^l\right)}{2} \right)^k \\
&\leqslant \frac{\alpha_{n-k}\alpha_k}{\alpha_n} (\text{Lip}\varphi)^k \liminf_{l\to\infty} \sum_j \left(\frac{\text{diam } B_j^l}{2} \right)^n \\
&\leqslant \frac{\alpha_{n-k}\alpha_k}{\alpha_n} (\text{Lip}\varphi)^k \mathcal{H}_n(S).
\end{aligned}$$

Next, we should verify that $S \cap \varphi^{-1}(y)$ is $\mathcal{H}_{n-k}$ measurable for $\mathcal{H}_k$ *a.e.* y, and $y \in \mathbb{R}^k \mapsto \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right)$ is $\mathcal{H}_k$ measurable. First, assuming that S is compact, we can write

$$\mathcal{H}^*_{n-k}\left(S \cap \varphi^{-1}(y)\right) = \lim_{\delta\downarrow 0} \mathcal{H}^*_{n-k,\delta}\left(S \cap \varphi^{-1}(y)\right) = \sup_{\delta>0} \mathcal{H}^*_{n-k,\delta}\left(S \cap \varphi^{-1}(y)\right).$$

Note that in this case, $S \cap \varphi^{-1}(y)$ is a Borel set and thus $\mathcal{H}_{n-k}$ measurable, therefore it suffices to verify that $y \mapsto \mathcal{H}^*_{n-k,\delta}\left(S \cap \varphi^{-1}(y)\right)$ is $\mathcal{H}_k$ measurable for all $\delta > 0$. Actually, $\mathcal{H}^*_{n-k,\delta}\left(S \cap \varphi^{-1}(y)\right)$

is upper semicontinuous. To see this, note that the spherical Hausdorff outer measure can also be defined by open balls, and thus for arbitrary $\epsilon > 0$, there exits an at most countable collection of open balls $\{B_j\}$ such that $S \cap \varphi^{-1} \subset \bigcup_j B_j$, for all j, diam $B_j \leqslant \delta$, and

$$\sum_j \alpha_{n-k} \left(\frac{\text{diam } B_j}{2}\right)^{n-k} \leqslant \mathcal{H}^*_{n-k,\delta}\left(S \cap \varphi^{-1}(y)\right) + \epsilon.$$

The compactness of S implies that if $|y - y'|$ small enough, then $S \cap \varphi^{-1}(y') \subset \bigcup_j B_j$. Therefore,

$$\limsup_{y' \to y} \mathcal{H}^*_{n-k,\delta}\left(S \cap \varphi^{-1}(y')\right) \leqslant \mathcal{H}^*_{n-k,\delta}\left(S \cap \varphi^{-1}(y')\right) + \epsilon, \tag{31}$$

then the upper semicontinuity follows from the arbitrariness of ϵ. Second, let S be just measurable, then by the regularity of Lebesgue measure, there exists a sequence of compact sets $C_1 \subset C_2 \subset \ldots$ such that $S \backslash \bigcup_{i=1}^{\infty} C_i$ is of zero Lebesgue measure. By the cube decomposition of open set and the Eilenberg inequality,

$$\int^*_{\mathbb{R}^k} \mathcal{H}^*_{n-k,\delta}\left(\left(S \backslash \bigcup_{i=1}^{\infty} C_i\right) \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y = 0,$$

i.e. $\mathcal{H}^*_{n-k,\delta}\left(\left(S \backslash \bigcup_{i=1}^{\infty} C_i \cap \varphi^{-1}(y)\right)\right) = 0$, $\mathcal{H}_k$ *a.e.* y. The *a.e.* measurability of $S \cap \varphi^{-1}(y)$ and measurability of $\mathcal{H}^*_{n-k}\left(S \cap \varphi^{-1}(y)\right)$ follow from

$$\mathcal{H}^*_{n-k}\left(S \cap \varphi^{-1}(y)\right) = \mathcal{H}^*_{n-k}\left(\bigcup_{i=1}^{\infty} C_i \cap \varphi^{-1}(y)\right) + \mathcal{H}^*_{n-k}\left(\left(S \backslash \bigcup_{i=1}^{\infty} C_i \cap \varphi^{-1}(y)\right)\right),$$

and now we can use $\mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right)$ instead of $\mathcal{H}^*_{n-k}\left(S \cap \varphi^{-1}(y)\right)$.

part 2 Assuming that $J\varphi(x) > 0$ for all $x \in E$. We use an estimates of $|\varphi(x) - \varphi(y)|$ similar to (27): Let $k \in \{1, 2, \ldots, n\}$, $E \subset \mathbb{R}^n$ be an open set, $\varphi : E \to \mathbb{R}^k$ be a C^1 function such that $J\varphi(x) > 0$ for all $x \in E$, then for arbitrary $x_0 \in E$ and for all $\epsilon > 0$, there exists $\delta > 0$, such that open ball $B(x_0, \delta) \subset E$ and

$$|\varphi(x) - \varphi(y) - \varphi'(x_0)(x - y)| \leqslant \epsilon|x - y| \tag{32}$$

for all $x, y \in B(x_0, \delta)$.

Let $x_0 \in E$, since $J\varphi(x_0) > 0$, without loss of generality, assuming that the first k columns $\left\{\frac{\partial}{\partial x_1}\varphi(x_0), \frac{\partial}{\partial x_2}\varphi(x_0), \ldots, \frac{\partial}{\partial x_k}\varphi(x_0)\right\}$ are linear independent. Define $\Phi(x) = (\varphi(x), x_{k+1}, \ldots, x_n)$, by the implicit function theorem, Φ is a bijection on a neighborhood of x_0. By definition,

$$\begin{aligned}\Phi(x) &= \left(\Phi(x) - \Phi'(x_0)(x - x_0)\right) + \Phi'(x_0)(x - x_0) \\ &= \Phi'(x_0)\left[\left(\Phi'(x_0)\right)^{-1}\left(\Phi(x) - \Phi'(x_0)(x - x_0)\right) + (x - x_0)\right],\end{aligned}$$

denote the term in square brackets on the right hand side of the second equality as $g(x)$, then by estimate (32), for all $\epsilon > 0$ there exists $\delta > 0$ such that

$$(1-\epsilon)|x-y| \leqslant |g(x) - g(y)| \leqslant (1+\epsilon)|x-y|, \tag{33}$$

for all $x, y \in B(x_0, \delta)$. Therefore, $\varphi = A \circ g$ on $B(x_0, \delta)$, where $A = P\Phi'(x_0)$, $P : \mathbb{R}^n \to \mathbb{R}^k$ is the coordinate projection of the first k dimensions. By (33), $1-\epsilon \leqslant |g'| \leqslant 1+\epsilon$, by definition,

$$(J\varphi(x))^2 = \det\left(\varphi'(x)\varphi'(x)^T\right) = \det\left(A \circ g'(x) \circ g'(x)^T \circ A^T\right).$$

By SVD / PD, $g'(x) \circ g'(x)^T = Q^T C Q$, where C is diagonal with $(1-\epsilon)^2 \leqslant c_{ii} \leqslant (1+\epsilon)^2$, $i \in \{1, 2, \ldots, n\}$, $Q^T Q = I_n$; $A = PU^T$, where $P \in \mathbb{R}^{k\times k}$ is symmetric and $U \in \mathbb{R}^{n\times k}$ is orthogonal. Therefore,

$$\det\left(A \circ g'(x) \circ g'(x)^T \circ A^T\right) = \det\left(PU^T Q^T CQUP^T\right) = (\det(A))^2 \det\left(U^T Q^T CQU\right)$$

Note that QU is also orthogonal, then

$$(1-\epsilon)^{2k} \leqslant \det\left(U^T Q^T CQU\right) \leqslant (1+\epsilon)^{2k}.$$

As a result,

$$(1-\epsilon)^k JA \leqslant J\varphi(x) \leqslant (1+\epsilon)^k JA \tag{34}$$

on $B(x_0, \delta)$. Compare (34) and (28), then see that the ideas of the **step** 4 and **step** 5 of the proof sketch of Proposition T.3.1 can be used here.

When $J\varphi(x) = 0$, let $\epsilon > 0$, define $\varphi_\epsilon : E \times \mathbb{R}^k \to \mathbb{R}^k$ as

$$(x, z) \mapsto \varphi(x) + \epsilon z,$$

$Crit(\varphi) = \{x \in E | J\varphi(x) = 0\}$. Note that $J\varphi_\epsilon(x, z) > 0$ for all $(x, z) \in E \times \mathbb{R}^k$, for arbitrary $w \in \mathbb{R}^k$,

$$\begin{aligned}\int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y &= \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y - \epsilon w)\right) d\mathcal{H}_k y \\ &= \frac{1}{\alpha_k} \int_{B(0,1)} \int_{\mathbb{R}^{n-k}} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y - \epsilon w)\right) d\mathcal{H}_k y d\mathcal{H}_k w.\end{aligned}$$

Let $P' : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^k$ be the coordinate projection of the last k dimensions, $S' = S \times B(0,1) \subset$

$\mathbb{R}^{n+k}$. Then note that for all $y \in \mathbb{R}^k$, $w \in B(0,1)$,

$$S' \cap \varphi_\epsilon^{-1}(y) \cap \left(P'\right)^{-1}(w) = \left(S \cap \varphi^{-1}(y - \epsilon w)\right) \times \{w\}.$$

Therefore, by the Eilenberg inequality and the Fubini-Tonelli theorem,

$$\begin{aligned}
&\int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y \\
&\quad = \frac{1}{\alpha_k} \int_{B(0,1)} \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S' \cap \varphi_\epsilon^{-1}(y) \cap \left(P'\right)^{-1}(w)\right) d\mathcal{H}_k y d\mathcal{H}_k w \\
&\quad \leqslant \frac{\alpha_{n-k}}{\alpha_n} \int_{\mathbb{R}^k} \left(\mathrm{Lip} P'\right)^k \mathcal{H}_n\left(S' \cap \varphi_\epsilon^{-1}(y)\right) d\mathcal{H}_k y \\
&\quad \leqslant \frac{\alpha_{n-k}}{\alpha_n} \int_{S'} J\varphi_\epsilon(x,z)\, d\mathcal{H}_{n+k}(x,z) \leqslant \frac{\alpha_{n-k}\alpha_k}{\alpha_n} \mathcal{L}_n(S) \sup_{(x,z)\in S'} J\varphi_\epsilon(x,z).
\end{aligned}$$

The last inequality above uses the fact that $\mathcal{H}_{n+k} = \mathcal{L}_{n+k}$ is the completion of $\mathcal{L}_n \times \mathcal{L}_k$. If E is bounded, the right hand side of the last inequality converges to 0 as $\epsilon \downarrow 0$ Then conclude for (not necessarily bounded) open set $E \subset \mathbb{R}^n$ by the cube decomposition. ■

Proof of Theorem 3.1.2.

We show that Theorem 3.1.2 follows from Proposition T.3.3, using the Rademacher-Whitney-Lusin-Egoroff framework as in **part** 2 of the proof sketch of the Theorem 3.1.1. Without loss of generality, let E be bounded. Suppose we already find a C^1 function $\overline{\varphi}$ and compact set C', such that (29) holds for an $\epsilon > 0$. Now, by the Eilenberg inequality,

$$\begin{aligned}
\int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(S \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y &= \int_{S\cap C'} J\varphi(x)\, d\mathcal{L}_n x + \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(\left(S \backslash C'\right) \cap \varphi^{-1}(y)\right) d\mathcal{H}_k y \\
&\leqslant \int_{S\cap C'} J\varphi(x)\, d\mathcal{L}_n x + \frac{\alpha_{n-k}\alpha_k}{\alpha_n} (\mathrm{Lip}\varphi)^k \mathcal{L}_n\left(S \backslash C'\right) \\
&\leqslant \int_{S\cap C'} J\varphi(x)\, d\mathcal{L}_n x + M_1 \epsilon,
\end{aligned}$$

where M_1 is a constant that does not depend on ϵ. Since φ is Lipschitz, there exists another constant M_2, such that $J\varphi < M_2$, and thus,

$$\int_{S\backslash C'} J\varphi(x)\, d\mathcal{L}_n x \leqslant M_2 \epsilon.$$

Then conclude by the same discussion as **part** 3 of the proof sketch Theorem 3.1.1. ■

Corollary T.3.4. *Another Sard type lemma*

Let $k \in \{1, 2, \ldots, n\}$, $E \subset \mathbb{R}^n$ be an open set, $\varphi : E \to \mathbb{R}^k$ be a C^1 function. Then for $\mathcal{L}_k$ a.e. $y \in$

$\mathbb{R}^k$,

$$\mathcal{H}_{n-k}\left(Crit\left(\varphi\right) \cap \varphi^{-1}\left(y\right)\right) = 0 \tag{35}$$

and $\varphi^{-1}\left(y\right)\backslash Crit(\varphi)$ *can be locally parameterized by implicit functions.*

PROOF. Since φ is a C^1 function, $J\varphi : E \to \mathbb{R}$ is continuous. Therefore, $Crit\left(\varphi\right) = \{x \in E : J\varphi\left(x\right) \geqslant 0\}\backslash\{x \in E : J\varphi\left(x\right) > 0\}$ is $\mathcal{L}_n$ measurable. Let $y \in \mathbb{R}^k$, by (8),

$$\int_{Crit(\varphi)} J\varphi\left(x\right) d\mathcal{L}_n x = \int_{\mathbb{R}^k} \mathcal{H}_{n-k}\left(Crit\left(\varphi\right) \cap \varphi^{-1}\left(y\right)\right) d\mathcal{H}_k y.$$

The integral in the left hand side of the above equality is 0, thus by the property of Lebesgue integral and the fact that $\mathcal{L}_k = \mathcal{H}_k$ in $\mathbb{R}^k$, (35) holds for $\mathcal{L}_k$ *a.e.* $y \in \mathbb{R}^k$.

For arbitrary $x \in \varphi^{-1}\left(y\right)\backslash Crit(\varphi)$ where y satisfies (35), $J\varphi\left(x\right) > 0$. Note that $\varphi'\left(x\right)$ is row full rank. Therefore, the implicit function theorem can be applied in a neighborhood of x. Specifically, pick k linear independent columns of $\varphi'\left(x\right)$, without loss of generality, assuming that the first k columns $\left\{\frac{\partial}{\partial x_1}\varphi\left(x\right), \frac{\partial}{\partial x_2}\varphi\left(x\right), \ldots, \frac{\partial}{\partial x_k}\varphi\left(x\right)\right\}$ are linear independent. Let U be a neighborhood of x, define

$$\Psi\left(x_1', \ldots, x_k', x_{k+1}', \ldots, x_n'\right) = \varphi\left(x_1', \ldots, x_k', x_{k+1}', \ldots, x_n'\right) - y,$$

where $y = \varphi\left(x\right)$, then $\nabla_{x_1, x_2, \ldots, x_k}\Psi\left(x_1, \ldots, x_k, x_{k+1}, \ldots, x_n\right)$ is full rank, and thus there exists a C^1 implicit function g in a open subset $B_{x_1,\ldots,x_k} \times B_{x_{k+1},\ldots,x_n} \subset U$, where

$$\begin{aligned} B_{x_1,\ldots,x_k} &= \left\{\left(x_1', \ldots, x_k'\right) \in \mathbb{R}^k : \left|\left(x_1', \ldots, x_k'\right) - \left(x_1, \ldots, x_k\right)\right| < \alpha\right\}, \\ B_{x_{k+1},\ldots,x_n} &= \left\{\left(x_{k+1}', \ldots, x_n'\right) \in \mathbb{R}^{n-k} : \left|\left(x_{k+1}', \ldots, x_n'\right) - \left(x_{k+1}, \ldots, x_n\right)\right| < \beta\right\}, \end{aligned}$$

such that $\Psi\left(x_1', \ldots, x_k', x_{k+1}', \ldots, x_n'\right) = 0 \Leftrightarrow \left(x_1', \ldots, x_k'\right) = g\left(x_{k+1}', \ldots, x_n'\right)$ for all $x' \in B_{x_1,\ldots,x_k} \times B_{x_{k+1},\ldots,x_n}$. Now, define

$$\psi\left(x_{k+1}', \ldots, x_n'\right) = \left(g\left(x_{k+1}', \ldots, x_n'\right), x_{k+1}', \ldots, x_n'\right)^T,$$

then ψ is a C^1 diffeomorphism and thus a homeomorphism. ∎